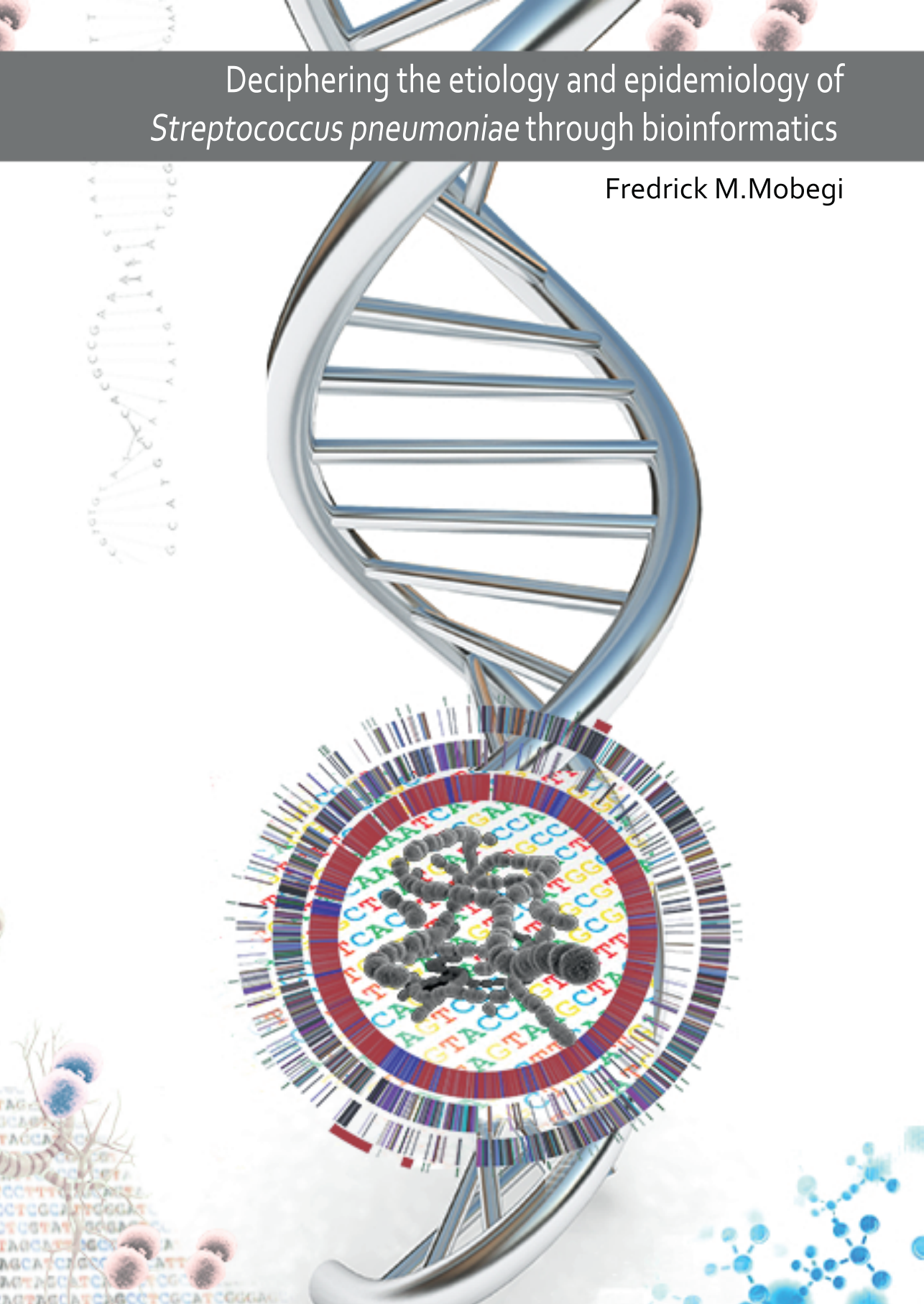
Deciphering the etiology and epidemiology of *Streptococcus pneumoniae* through bioinformatics

Fredrick M.Mobegi



Deciphering the etiology and epidemiology of *Streptococcus pneumoniae* through bioinformatics

Fredrick M. Mobegi

The research presented in this thesis was conducted at the Laboratory of Pediatrics infectious Diseases, and the Bacterial Genomics group at the Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands.

Printing of this thesis was financially supported by the Radboud University Nijmegen, the Netherlands.

Cover design:	F.M. Mobegi
Layout:	F.M. Mobegi
Printing:	GVO, Ede
ISBN:	978-94-6332-096-2

Deciphering the etiology and epidemiology of *Streptococcus pneumoniae* through bioinformatics

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 20 december 2016
om 11.30 uur precies

door

Fredrick Maati Mobegi

geboren op 1 oktober 1984
Kisii North, Kenya

Promotor:

Prof. dr. P.W.M. Hermans
Prof. dr. R. de Groot

Copromotoren:

Dr. A.L. Zomer (Universiteit Utrecht)
Dr. S.A.F.T. van Hijum

Manuscriptcommissie:

Prof. dr. H.F. Wertheim
Prof. dr. M.A. Huijnen
Prof. dr. J.M. van Dijl (UMC Groningen)

Deciphering the etiology and epidemiology of *Streptococcus pneumoniae* through bioinformatics

Doctoral thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M van Krieken,
according to the decision of the Council of Deans
to be defended in public on

Tuesday, 20 December 2016
at 11:30 hours

by

Fredrick Maati Mobegi

born on Monday, 1 October 1984
in Kisii North, Kenya.

Supervisors:

Prof. dr. P.W.M. Hermans
Prof. dr. R. de Groot

Co-supervisors:

Dr. A.L. Zomer (Utrecht University)
Dr. S.A.F.T. van Hijum

Manuscript committee:

Prof. dr. H.F. Wertheim
Prof. dr. M.A. Huijnen
Prof. dr. J.M. van Dijl (UMC Groningen)

To my wife and my family: your words of encouragement and push for tenacity always kept me going.

We can only be said to be alive in those moments when our hearts are conscious of our treasures. Thornton Wilder (1897-1975)

Table of contents

Chapter 1	General introduction	11
Chapter 2	Advances and perspectives in computational prediction of microbial gene essentiality	31
Chapter 3	From microbial gene essentiality to novel antimicrobial drug targets	53
Chapter 4	Post-vaccine microevolution of invasive *Streptococcus pneumoniae*	73
Chapter 5	Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data	89
Chapter 6	Genetic microbial correlates of the clinical manifestation of invasive pneumococcal disease	119
Chapter 7	Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease	139
Chapter 8	General discussion	155
Addendum		177
	Dutch summary (Samenvatting)	179
	Author affiliation	183
	Acknowledgement	184
	List of publications	186
	Other activities	188
	Curriculum vitae	189

Chapter 1

General introduction

Background

Streptococcus pneumoniae: a concise history

Streptococcus pneumoniae, or the pneumococcus, is a Gram-positive, catalase-negative, alpha-hemolytic facultative anaerobic bacterium. Pneumococcus belongs to the phylum Firmicutes, class Bacilli, order Lactobacillales, family *Streptococcaceae*, genus *Streptococcus*, and species *S. pneumoniae*. The pneumococcus was first identified in 1881 when two microbiologists, George Miller Sternberg - a physician in the United States Army, and Louis Pasteur - a French chemist, simultaneously and independently isolated and described a roughly lancet-shaped pair of coccoid bacteria in human saliva. In their experiments, they injected rabbits with saliva from human carriers and subsequently recovered diplococci from the blood of these rabbits [1, 2]. The infected rabbits soon died from septicemia revealing the pathogenic potential of the newly discovered bacterium in mammals. They described a microorganism with characteristic paired cell, diplococcal morphology and a distinct capsular layer around each bacterium. Since these discoveries, antibiotics and vaccines have been discovered to combat the pneumococcus, but still, infection by *S. pneumoniae* cause up to 1.6 million mortalities globally each year [3].

Identification and characterization of the pneumococcus

Laboratory identification of pneumococci and routine bacteriology continue to rely on decades-old culture-based methods [4]. Diagnosis of pneumococci etiology relies on recovery, usually verified by a positive culture, of the organism from patients' diagnostics samples (sputum, lungs, spinal fluid, or blood) [5]. *S. pneumoniae* grows best at 35-37°C with ~5% CO_2 on blood agar but can also grow on a chocolate agar. Subtle differences between pneumococci and other genetically related strains and viridans streptococci can be used for accurate diagnostics. *S. pneumoniae* can be identified using Gram's stain, catalase, and optochin tests simultaneously, and the results confirmed using a bile solubility test. If these tests indicate the presence of *S. pneumoniae*, serological or DNA-based molecular typing approaches can then be performed to identify the serotype [6-8]. This order of analysis, as summarized in Figure 1, is an efficient way to save on time and costly serotyping chemicals. It is noteworthy that pneumococcal capsular serotyping is not usually required for a clinical response but, it is a crucial part of successful pneumococcal disease surveillance efforts especially in the light of pneumococcal conjugate vaccines, which targets the highly variable capsule.

Catalase test: Catalase breaks down hydrogen peroxide (H_2O_2) into water (H_2O) and oxygen (O_2). In catalase-positive microorganisms, the oxygen is exuded as bubbles in the liquid, a phenomenon lacking in catalase-negative microorganisms [9]. The catalase test is mainly used to distinguish between Gram-positive cocci. Members of the genera *Streptococcus* and *Enterococcus* are catalase-negative while members of the genus *Staphylococcus* are catalase-positive.

Gram's stain: The method is named after its inventor, Hans Christian Gram, a Danish scientist who was experimenting with the techniques for visualizing bacteria in Carl Friedländer's laboratory [10]. By sequentially exposing sections of lung tissues to aniline-gentian violet; a weak solution of ethanol, iodine, and Bismarck brown or vesuvin, Gram observed many pairs of slightly elongated cocci that retained the dark aniline-gentian violet stain. Other bacteria in Gram's specimens failed to retain the aniline-gentian violet stain. Gram's phenomenal discovery [11] that virtually all clinically relevant bacteria are either Gram-positive or Gram-negative became one of the foundations of clinical microbiology.

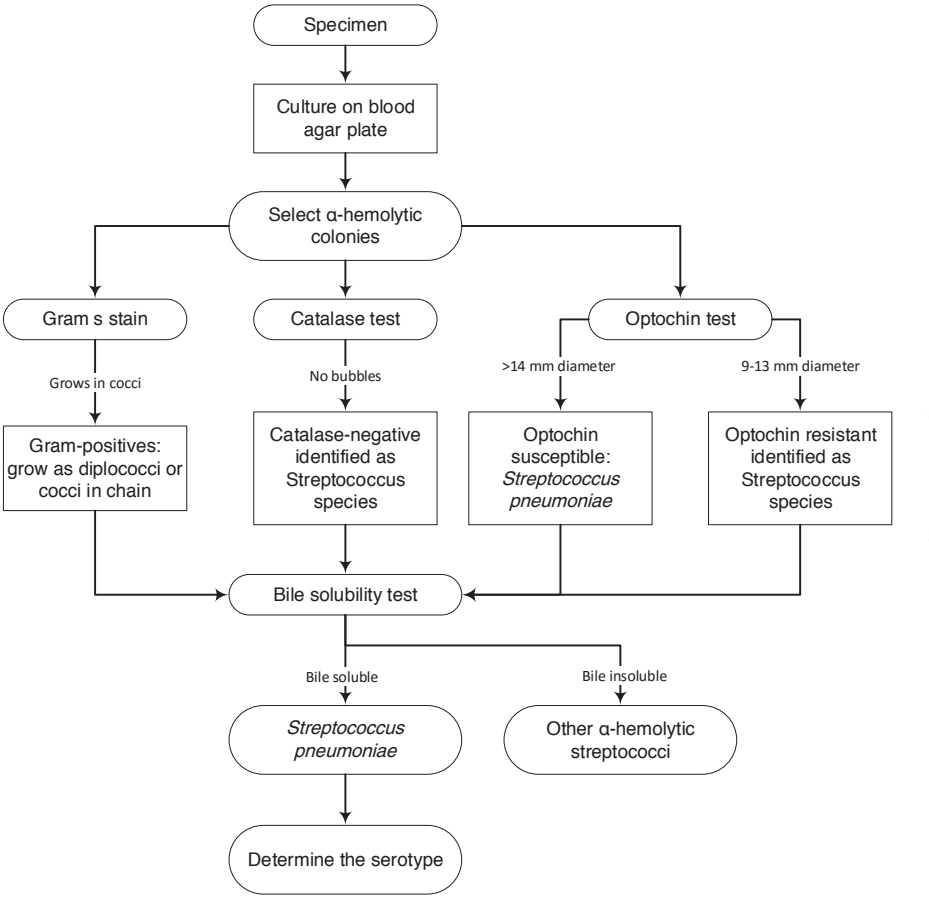


Figure 1. Schematic overview of the steps for identifying and characterizing *S. pneumoniae*

Bile solubility test and Optochin Test: When grown aerobically on blood agar, *S. pneumoniae* like many other streptococci produce hydrogen peroxide which induces alpha (α)-hemolysis by oxidizing hemoglobin to green methemoglobin; the characteristic green coloration observed on blood agar plates [12]. A distinguishing feature between *S. pneumoniae* and other α-hemolytic streptococci is its sensitivity to lysis by bile, in the so-called 'bile solubility test' [13]. The bile salt deoxycholate activates pneumococcal LytA amidase, an essential autolytic enzyme [14]. Alternatively, *S. pneumoniae* can be differentiated based on its susceptibility to optochin (ethylhydrocupreine); optochin inhibits the pneumococcal F_0F_1-ATPase [15]. However, none of these tests are entirely accurate because resistance of *S. pneumoniae* to both deoxycholate [16] and optochin [17] has been observed.

Serology: Quellung test [8], an immunological reaction with type-specific pneumococcal antisera is the gold standard microbiological analysis for identification and typing of *S. pneumoniae*. However, serological tests cannot be used to identify pneumococcal strains lacking a polysaccharide capsule. Cross-reaction may also occur with viridans streptococci which have an antigen similar to pneumococcal polysaccharide leading to false-positive results. Furthermore, use of older colonies or mixed cultures may lead to equivocal serological reactions.

Polymerase chain reaction (PCR): PCR [18] is a valuable tool for selective amplification of particular target DNA sequence(s) within a heterogeneous collection of DNA sequences. Conventional and real-time PCR assays have been used for diagnosis or surveillance of pneumococci by detecting specific sequence regions of the pneumolysin (*ply*), autolysin (*lytA*), and pneumococcal surface adhesion (*psaA*) genes [19, 20]. Several pneumococcal serospecificities are included in multiplex PCR schemes [6, 7, 21]. Furthermore, PCR can be combed with partial or targeted genome sequencing [22] or transcriptomics [23] to improve the detection of *S. pneumoniae*. Nonetheless, discrepant results, although rare, may occur in PCR-based assays due to recombination events that occur between pneumococci and closely related and viridans streptococci [20, 21]. Furthermore, viridans streptococci may sometimes also contain the target genes [24].

Molecular typing using MLST: Multilocus sequence typing; MLST [25] relies on amplification and sequencing of up to seven housekeeping genes (Shikimate dehydrogenase; *aroE*, glucose-6-phosphate dehydrogenase; *gdh*, glucose kinase; *gki*, xanthine phosphoribosyltransferase; *xpt*, transketolase; *recP*, Signal peptidase I; *spi*, and D-alanine-D-alanine ligase; *ddl*), and using the allelic variation in these genes to assign the pneumococcal isolates to sequence types or clones [26]. The validity of MLST is based on the paradigm that isolates with identical or closely related allelic profiles typically express the same serotype [26], but few exceptions have been reported [27]. MLST is widely used, mainly for epidemiological surveys and population genomics, because it offers a higher resolution to distinguish isolates at a species level compared to other techniques such as 16S RNA sequencing [27, 28], and is much cheaper than whole-

genome sequencing (WGS) which provides superior resolution. Unfortunately, MLST has little power in resolving small evolutionary differences and thus not suitable for characterizing possible differences in strains within pneumococcal outbreaks. To improve the discriminatory power, MLST data can be used in combination with pulsed-field gel electrophoresis (PFGE), Multiple-locus variable number tandem repeat analysis (MLVA) [29], or sequencing data from more variable genetic loci such as penicillin binding protein (pbps) and pneumococcal surface adhesin A (*psaA*). It will only be a matter of time before the robustness of WGS could be put to routine clinical use in the diagnosis of the pneumococcus. In fact, WGS is already pivotal in pneumococcal molecular epidemiology studies [30], key among them the Global Pneumococcal Sequencing Project [31]. This global database will provide a platform to measure the response of *S. pneumoniae* to vaccines and precisely characterize patterns of genetic adaptation on global and regional levels, an insight that will significantly inform the designing of future interventions.

Pneumococcal nasopharyngeal colonization

Although asymptomatic carriage is common in healthy individuals, nasopharyngeal colonization is a prerequisite for pneumococcal disease and a source of the pneumococcal spread between individuals [32, 33]. The pneumococci are spread through contact with ailing or healthy carriers of the bacteria. Transmission from person to person occurs via respiratory droplets/aerosols from the nose or mouth. Children acquire pneumococcal approximately twice as much as adults do [34]. The density and duration of pneumococcal colonization are mostly influenced by the host age [35] and serotype of the colonizing strain [36]. Certain serotypes including type 6B, 9V, 14, 19F and 23F are most commonly found in young children [4] whereas type 1, 3, 5, 7F/A, 14 and 19A are most commonly found in adults [37, 38]. Overall, common pneumococcal serotypes have been observed to be superior in acquisition rate and carriage duration, and they are generally non-susceptible to colonization competition [36]. Pneumococcal carriage is usually high in children with disproportionately higher rates in developing countries compared developed countries [39]. For most serotypes, the prevalence of nasopharyngeal colonization peaks around 2-3 three years of life and diminishes after that to <10% in adults [4]. The duration of carriage also declines with age [36], probably as a result of the mature immunologic response which develops with age. Interestingly, colonization remains low in old age > 65 years despite the waning immunity [40]. Children are therefore the primary reservoir of pneumococcal strains that circulate within communities. The colonization rates in young children associate with the high risk of invasive pneumococcal diseases (IPD). However, the incidence of IPD observed in the elderly is remarkably high despite the low carriage rates, probably due to the heterogeneity in disease susceptibility among the elderly [32].

Apart from age-dependent colonization, other lifestyle and dietary factors, including living in crowded area, or in close-contact like a day care center [33], consumption of, or exposure to tobacco [41], and excessive sugar intake in early age [42] are also implicated

in promoting the acquisition of, and infection by *S. pneumoniae*. Although probably linked to the underlying socioeconomic factors which also affect colonization rates, studies have shown that host genetic factors may influence pneumococcal colonization in various ethnicities [43, 44].

Epidemiology of pneumococcal disease

The pneumococcus has a dichotomous lifestyle, persisting as a commensal colonizer of the nasopharynx in healthy carriers, but can also transition into a ferocious opportunistic pathogen in susceptible individuals. Pneumococcal infections mainly consist of superficial mucosal infections, such as sinusitis, conjunctivitis, acute otitis media, acute exacerbations of chronic bronchitis, and exacerbation of chronic obstructive pulmonary disease (COPD). However, when the organism disseminates into deeper, usually sterile innate tissue compartments (Figure 2), it can instigate severe invasive pneumococcal diseases (IPD), including pneumonia, septic bacteremia, and meningitis [45]. IPD is associated with a high degree of morbidity and mortality worldwide.

Age and gender are important risk factors for pneumococcal infections. Disease risk is highest in young children <2 years, the aged >65 years, and individuals with chronic illnesses or immune suppression [46]. The incidence of IPD is up to 50 times higher in the young (under 5 years old) and the elderly (over 65 years old) than in adolescents [47]. Additionally, there is a predominance of pneumococcal disease in males than female [47]. This may be because underlying conditions such as cigarette smoking and alcoholism are more prevalent among men. An increased risk of IPD is also concomitant with defects in the nonspecific or specific defense mechanisms against colonization or invasion of *S. pneumoniae*. Other important risk factors include chronic and immunocompromising illnesses such as heart, lung and kidney disease, sickle cell disease, diabetes, cancer, HIV/AIDS asplenia, and admission into group childcare or nursing homes [48, 49]. Approximately 1.6 million people, including up to a million children under five years old, die of IPD every year. IPD is the leading cause of death among vaccine-preventable infectious diseases [3]. Developing countries bear the greatest burden of IPD than industrialized countries [45], probably due to extensive immunization campaigns and socioeconomic status [50].

S. pneumoniae is the most common cause of community-acquired pneumonia worldwide [51], and together with *H. influenzae* and *M. catarrhalis*, they are the major bacterial causes of respiratory tract infections [52]. Secondary bacterial infections are thought to be the cause majority of the deaths in viral pandemics. Pneumococcal co-infections are particularly common in viral pandemics, especially those involving influenza virus [53]. Pneumococcal pneumonia is characterized by four stages. First, the lung alveoli are filled up with fluids that contain pneumococci. These fluids mediate the spread of the organism throughout the lungs. The host immunity recruits neutrophils and red blood cells which invade the alveoli inducing an inflammatory immune response. By this stage, most of the

bacteria have been cleared, and the neutrophils packed into the alveoli. Finally, macrophages eliminate the remaining residue from the inflammatory response effectively dissipating the assault [54]. Much of the damage to the lungs is therefore mainly caused by the host's inflammatory response.

If it manages to persist in the lungs, the pneumococcus can invade the blood and cause bacteremia which could lead to sepsis [46]. When in the blood it can cross the blood-brain barrier and infect the meninges causing meningitis [55]. The pneumococcus is also implicated in causing milder infections in other parts, including the middle ear; leading to otitis media [56], the paranasal sinuses; causing sinusitis [57], the peritoneum; causing peritonitis [58], and the joints or bones; causing septic arthritis or osteomyelitis respectively [59]. The polysaccharide capsule is thought to be the major virulent factor and the primary determinant of the pneumococcal invasive potential [60]. Recent studies, however, suggest that other determinants of invasiveness may be localized beneath the capsule [61-63]. These factors should be included in evaluating the effects of current vaccines and modeling future interventions [64].

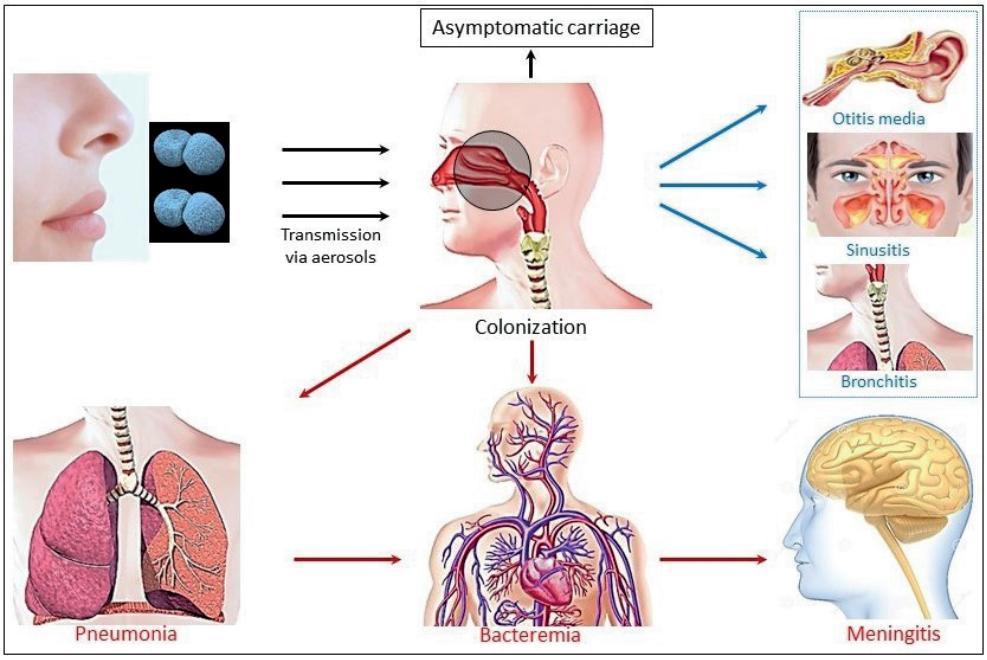


Figure 2. Overview of pneumococcal colonization, transmission and progression to disease. Adapted from Henriques-Normark and Tuomanen (2013)

Prevention and treatment of pneumococcal disease

There are several different prevention and therapeutic options for pneumococcal infections. For the prevention of pneumococcal disease, two types of vaccines which contain purified capsular polysaccharides of the most common serotypes that cause IPD are currently offered. These are pneumococcal polysaccharide vaccine; PPV, and pneumococcal conjugate vaccine; PCV [65]. PPV primarily induces a B-cell-dependent immune response and require a revaccination after 5-6 years. They are therefore recommended for use in adults, because they have a matured immune system, but not for children under the age of two [50, 65]. PCV was originally available as a *heptavalent*, containing 7 serotypes (PCV7), but was later extended to include *decavalent* (PCV10) and *tridecavalent* (PCV13) prescriptions. These PCVs are used for childhood immunization programs in most countries around the world. Pneumococcal vaccines have significantly reduced episodes of pneumococcal infections caused by protected serotypes, also known as vaccine-type serotypes, in children and non-vaccinated adults signifying herd-immunity [45]. But due to the limited coverage by the PCV (protecting against 7, 10, or 13 serotypes), and the considerable antigenic variability in *S. pneumoniae* capsular types; with over 90 antigenically distinct serotypes characterized to date [22, 23], a universal vaccine has not been developed. 'Serotype replacement' with non-vaccine serotypes also occurs rapidly under vaccine pressure [30, 66]. The replacement, at least partially, abrogates the protective effect of immunization. In rare cases, strains may exchange entire capsular loci, in the process commonly referred to as 'capsule switching', leading to new genotypes capable of evading vaccines or exhibiting resistance to essential antibiotic [30, 66].

Treatment of pneumococcal infections typically depends on the type of infection and disease severity. Routine use of antibiotics to treat minor bacterial infections is not recommended as it increases the probability of the bacteria developing drug-resistance over time [67]. Mild non-invasive pneumococcal infections such as sinusitis and bronchitis, are usually cleared by the host immunity without the need for treatment. Antibiotics could be used in rare cases for outpatient treatment or prophylaxis in patients with recurrent infections [68], especially those with underlying immunocompromising conditions such as HIV/AIDS and sickle cell anemia [69]. For IPD, however, antibiotic therapy - primarily with β-lactams - in inpatient medical ward or intensive care units is recommended [50, 70]. Nevertheless, the fact that *S. pneumoniae* is increasingly developing resistance to β-lactams [71] and to most other essential antibiotics [52, 70] is a greater concern to healthcare management systems globally. *S. pneumoniae* has a natural transformation system that facilitates the exchange of genetic material [67]. Therefore, strains that have developed antibiotic resistance can often spread these traits to other strains accelerating the problem. Due to the natural attributes of the pneumococcus in evading current remedies, proper management of the pneumococcus is a growing concern. Therefore, new tools for investigating the pneumococcus and developing potent therapies are required.

Pneumococcal genomics

Since its discovery in the 1880s, the pneumococcus has been extensively studied and was involved in several historical findings such as Gram staining, serotherapy, the discovery of DNA as the hereditary genetic material, and that bacteria can transfer genetic material via transformation [72, 73]. Complete sequencing of pneumococcal genomes was first completed in 2001 [74]. The genome contains a core set of around 1,500 genes that are essential for colonization, or growth and viability [30, 75, 76]. In the accessory or flexible genome, considerable plasticity exists between pneumococcal strains; mainly from insertional sequences which account for about 5% of the genome, and replete direct-repeat DNA elements and mobile (integrative and conjugative) genetic elements which provide recombinational hotspots for genetic variability [74, 75, 77]. Croucher, *et al.* for example, demonstrated that short-term pneumococcal variations are characterized by movement of phages and intragenomic rearrangements, with the slower transfer of stable loci distinguishing lineages [75]. The readiness to quickly take up genetic material enables the pneumococcus to gain and spread DNA sequences responsible for antibiotic resistance and capsular switches that may explain vaccine evasion [30, 66]. Pneumococci also encode a plethora of transport systems which, among other functions, help in nutrient acquisition, importing carbon and amino acid substrates, and in exporting adhesins, degradation enzymes, and components for capsular synthesis. These systems are also essential for the genetic competence of the pneumococcus and as drug-efflux pumps [74]. About 150 genes of the complement genome contribute to virulence and another set of about 170 genes actively maintain the non-invasive phenotype [50, 74], as invasion is considered an evolutionary dead-end for the microorganism. In other words, invasion does not evolutionarily benefit the pneumococcus since invading strains cannot transmit anymore.

In the recent past, high-throughput genome sequencing technologies (HTGST) have emerged as *de facto* tools that offer unparalleled resolution for microbiology. The technologies have significantly contributed to the expansion of dedicated databases such as GenBank that collate genome sequencing data. Using bioinformatics and genomics tools, information in these databases could be mined and applied to study molecular epidemiology, comparative genomics, functional genomics, and population genomics for the pneumococcus. For example, the technology has been used in combination with other *in silico* techniques such as homology mapping, comparative genomics, Tn-seq [78], and machine learning among others – as reviewed in **Chapter 2** - to predict essential genes which form potential drug targets [79], discover genetic variants underlying relevant phenotypes of antibiotic-resistance [80] and clinical manifestation of IPD [76], and to study the molecular epidemiology and evolution of pneumococcal strains [30, 75, 77]. Bentley and colleagues also utilized HTGST and comparative genomics to identify capsular biosynthesis gene clusters for 90 pneumococcal serotypes and explore how capsular biosynthesis and the evolution of capsular genes give rise to antigenic diversity in the pneumococcus [22]. These studies have advanced the general bacteriology of the

pneumococcus and provided invaluable knowledge that could steer the design of future interventions. Most importantly, they have accelerated the much-anticipated adoption of HTGST into routine applications in clinical laboratory microbiology and public health surveillance.

Research aims and outline of this thesis

The goal of the research described in this thesis is to understand better the etiology and molecular epidemiology of *Streptococcus pneumoniae* using bioinformatics tools. *In silico* bioinformatics studies were complemented with *in vitro* and *in vivo* experiments to generate novel insights into the etiology and molecular epidemiology of the pneumococcus. The rise of antibiotic-resistant bacteria has led to an urgent need for the development of novel diagnostics, drugs and vaccines, rapid detection of drug resistance in clinical isolates, and improvements in global surveillance. Given the importance of essential bacterial genes in cellular viability, they form promising drug targets for rational drug discovery. In **Chapter 2**, we review current advances and future perspective into the discovery of essential genes using computational approaches. **Chapter 3** describes a proof of concept study that combines high-density transposon mutagenesis, genome sequencing and integrative genomics for the identification of potential drug targets in *S. pneumoniae*, *Haemophilus influenzae,* and *Moraxella catarrhalis*. The chapter explores the exhaustive power of transposon insertions sequencing (Tn-seq) complemented with *in silico* tracking to rapidly identify essential genes. Candidate essential genes are further subject to comparative selection criteria to identify a subset that is nonconserved in the human host and absent in the commensal human gut microbiota. Four new targets in this subset were experimentally validated.

The undesirable trends of post-vaccine 'serotype replacement' and 'capsule switching' which repeal the protective effects of the vaccines have been reported in pediatric carriage. **Chapter 4** presents the impact of pneumococcal vaccination on whole genome epidemiology of adulthood IPD using whole-genome sequencing of 350 pneumococcal isolates from Nijmegen. In this chapter, we have established the effect of PCV7 on the pneumococcal population epidemiology and present HTGST as a powerful tool for the surveillance of circulating and infecting pneumococcal strains. In **Chapter 5** and **Chapter 6**, we further explore the use of HTGST and genome-wide association studies (GWAS) to identify the genetic variants underlying bacterial phenotypes of antibiotics resistance and clinical manifestation of invasive disease respectively. **Chapter 5** describes an approach that uses GWAS of 1680 systematically selected literature and in-house *S. pneumoniae* isolates to compute the 'distance to antibiotic resistance' using the cumulative effect of antibiotic resistance-conferring SNPs. In **Chapter 6**, we analyze 349 pneumococcal genomes to discover the genetic variants that correlate with clinical manifestation of IPD. Some patients have an increased risk of death after the invasion by *S. pneumoniae.* A phage gene that encodes a platelet binding protein, PblB, was identified as one of the candidate correlates of all-cause mortality within the first 30 days of hospitalization (30 day mortality). **Chapter 7** reports the biological mechanism behind this association. These studies highlight the efficiency of genome sequencing and integrative genomics in pneumococcal bacteriology. The thesis concludes (**Chapter 8**) with a discussion of the results of all our studies (**Chapters 3-7**). The insights gained are integrated with the

current state of literature to help place our findings in the light of healthcare application and provided a primer for future research on the pneumococcus.

References

1. Pasteur, L. Note sur une maladie nouvelle provoquée par la salive d'un enfant mort de la rage. *Bulletin de l'Academie de Medicine (Paris)* **series 21,** 94-103 (1881).
2. Sternberg, G.M. A fatal form of septicaemia in the rabbit, produced by subcutaneous injection of human saliva. An experimental research. *Nat Board Health Bull*, 781–783 (1881).
3. World Health Organization. Challenges in global immunization and the Global Immunization Vision and Strategy 2006-2015. *Wkly Epidemiol Rec* **81,** 190-195 (2006).
4. Henriques-Normark, B. & Tuomanen, E.I. The Pneumococcus: Epidemiology, Microbiology, and Pathogenesis. *Cold Spring Harbor Perspectives in Medicine* **3,** a010215 (2013).
5. Reller, L.B., Weinstein, M.P., Werno, A.M. & Murdoch, D.R. Laboratory Diagnosis of Invasive Pneumococcal Disease. *Clinical Infectious Diseases* **46,** 926-932 (2008).
6. Siira, L., Kaijalainen, T., Lambertsen, L., Nahm, M.H., Toropainen, M. & Virolainen, A. From Quellung to Multiplex PCR, and Back When Needed, in Pneumococcal Serotyping. *Journal of Clinical Microbiology* **50,** 2727-2731 (2012).
7. Brito, D.A., Ramirez, M. & de Lencastre, H. Serotyping *Streptococcus pneumoniae* by Multiplex PCR. *Journal of Clinical Microbiology* **41,** 2378-2384 (2003).
8. Habib, M., Porter, B.D. & Satzke, C. Capsular serotyping of *Streptococcus pneumoniae* using the Quellung reaction. *J Vis Exp*, e51208 (2014).
9. Chelikani, P., Fita, I. & Loewen, P.C. Diversity of structures and properties among catalases. *Cellular and Molecular Life Sciences CMLS* **61,** 192-208 (2004).
10. Austrian, R. The Gram stain and the etiology of lobar pneumonia, an historical note. *Bacteriol Rev* **24,** 261-5 (1960).
11. Gram, H.C. Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. *Fortschritte der Medizin (in German)*, 185-189 (1884).
12. Facklam, R. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. *Clinical Microbiology Reviews* **15,** 613-630 (2002).
13. Neufeld, F. Ueber eine specifische bakteriolytische Wirkung der Galle. *Zeitschrift für Hygiene und Infektionskrankheiten* **34,** 454-464 (1900).
14. Mosser, J.L. & Tomasz, A. Choline-containing Teichoic Acid As a Structural Component of Pneumococcal Cell Wall and Its Role in Sensitivity to Lysis by an Autolytic Enzyme. *Journal of Biological Chemistry* **245,** 287-298 (1970).
15. Fenoll, A., Muñoz, R., Garcia, E. & de la Campa, A.G. Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the Fo complex of the *Streptococcus pneumoniae Streptococcus oralis* H+-ATPases. *Molecular Microbiology* **12,** 587-598 (1994).
16. Obregón, V., García, P., García, E., Fenoll, A., López, R. & García, J.L. Molecular Peculiarities of the *lytA* Gene Isolated from Clinical Pneumococcal Strains That Are Bile Insoluble. *Journal of Clinical Microbiology* **40,** 2545-2554 (2002).
17. Robson, R.L., Essengue, S., Reed, N.A. & Horvat, R.T. Optochin resistance in *Streptococcus pneumoniae* induced by frozen storage in glycerol. *Diagn Microbiol Infect Dis* **58,** 185-90 (2007).
18. Mullis, K.B. & Faloona, F.A. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. in *Methods in Enzymology*, Vol. Volume 155 335-350 (Academic Press, 1987).

19. Carvalho, M.d.G.S., Tondella, M.L., McCaustland, K., Weidlich, L., McGee, L., Mayer, L.W., Steigerwalt, A., Whaley, M., Facklam, R.R., Fields, B., Carlone, G., Ades, E.W., Dagan, R. & Sampson, J.S. Evaluation and Improvement of Real-Time PCR Assays Targeting lytA, ply, and psaA Genes for Detection of Pneumococcal DNA. *Journal of Clinical Microbiology* **45,** 2460-2466 (2007).
20. Suzuki, N., Yuyama, M., Maeda, S., Ogawa, H., Mashiko, K. & Kiyoura, Y. Genotypic identification of presumptive *Streptococcus pneumoniae* by PCR using four genes highly specific for *S. pneumoniae*. *Journal of Medical Microbiology* **55,** 709-714 (2006).
21. Moreno, J., Hernández, E., Sanabria, O. & Castañeda, E. Detection and Serotyping of *Streptococcus pneumoniae* from Nasopharyngeal Samples by PCR-Based Multiplex Assay. *Journal of Clinical Microbiology* **43,** 6152-6154 (2005).
22. Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrell, B., Reeves, P.R., Parkhill, J. & Spratt, B.G. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2,** e31 (2006).
23. Turner, P., Hinds, J., Turner, C., Jankhot, A., Gould, K., Bentley, S.D., Nosten, F. & Goldblatt, D. Improved Detection of Nasopharyngeal Co-colonization by Multiple Pneumococcal Serotypes by Use of Latex Agglutination or Molecular Serotyping by Microarray. *Journal of Clinical Microbiology* **49,** 1784-1789 (2011).
24. Whatmore, A.M., Efstratiou, A., Pickerill, A.P., Broughton, K., Woodard, G., Sturgeon, D., George, R. & Dowson, C.G. Genetic Relationships between Clinical Isolates of *Streptococcus pneumoniae*, Streptococcus oralis, and *Streptococcus mitis*: Characterization of "Atypical" Pneumococci and Organisms Allied to S. mitis Harboring *S. pneumoniae* Virulence Factor-Encoding Genes. *Infection and Immunity* **68,** 1374-1382 (2000).
25. Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. & Spratt, B.G. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 3140-3145 (1998).
26. Enright, M.C. & Spratt, B.G. Multilocus sequence typing. *Trends in Microbiology* **7,** 482-487 (1999).
27. Coffey, T.J., Enright, M.C., Daniels, M., Morona, J.K., Morona, R., Hryniewicz, W., Paton, J.C. & Spratt, B.G. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. *Molecular Microbiology* **27,** 73-83 (1998).
28. Enright, M.C. & Spratt, B.G. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* **144,** 3049-3060 (1998).
29. Schouls, L.M., van der Ende, A., Damen, M. & van de Pol, I. Multiple-Locus Variable-Number Tandem Repeat Analysis of Neisseria meningitidis Yields Groupings Similar to Those Obtained by Multilocus Sequence Typing. *Journal of Clinical Microbiology* **44,** 1509-1518 (2006).
30. Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P. & Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45,** 656-63 (2013).
31. GPS partners. Global Pneumococcal Sequencing Project. Vol. 2016 (2016).

32. Simell, B., Auranen, K., Käyhty, H., Goldblatt, D., Dagan, R. & O'Brien, K.L. The fundamental link between pneumococcal carriage and disease. *Expert Review of Vaccines* **11**, 841-855 (2012).
33. Bogaert, D., van Belkum, A., Sluijter, M., Luijendijk, A., de Groot, R., Rumke, H.C., Verbrugh, H.A. & Hermans, P.W. Colonisation by *Streptococcus pneumoniae* and *Staphylococcus aureus* in healthy children. *Lancet* **363**, 1871-2 (2004).
34. Mosser, J.F., Grant, L.R., Millar, E.V., Weatherholtz, R.C., Jackson, D.M., Beall, B., Craig, M.J., Reid, R., Santosham, M. & O'Brien, K.L. Nasopharyngeal Carriage and Transmission of *Streptococcus pneumoniae* in American Indian Households after a Decade of Pneumococcal Conjugate Vaccine Use. *PLoS ONE* **9**, e79578 (2014).
35. Turner, P., Turner, C., Jankhot, A., Helen, N., Lee, S.J., Day, N.P., White, N.J., Nosten, F. & Goldblatt, D. A Longitudinal Study of *Streptococcus pneumoniae* Carriage in a Cohort of Infants and Their Mothers on the Thailand-Myanmar Border. *PLoS ONE* **7**, e38271 (2012).
36. Lipsitch, M., Abdullahi, O., D'Amour, A., Xie, W., Weinberger, D.M., Tchetgen, E.T. & Scott, J.A.G. Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model. *Epidemiology (Cambridge, Mass.)* **23**, 510-519 (2012).
37. Sherwin, R.L., Gray, S., Alexander, R., McGovern, P.C., Graepel, J., Pride, M.W., Purdy, J., Paradiso, P. & File, T.M., Jr. Distribution of 13-valent pneumococcal conjugate vaccine *Streptococcus pneumoniae* serotypes in US adults aged >/=50 years with community-acquired pneumonia. *J Infect Dis* **208**, 1813-20 (2013).
38. Ardanuy, C., Marimon, J.M., Calatayud, L., Gimenez, M., Alonso, M., Grau, I., Pallares, R., Perez-Trallero, E. & Linares, J. Epidemiology of invasive pneumococcal disease in older people in Spain (2007-2009): implications for future vaccination strategies. *PLoS One* **7**, e43619 (2012).
39. García-Rodríguez, J.Á. & Fresnadillo Martínez, M.J. Dynamics of nasopharyngeal colonization by potential respiratory pathogens. *Journal of Antimicrobial Chemotherapy* **50**, 59-74 (2002).
40. Palmu, A.A., Kaijalainen, T., Saukkoriipi, A., Leinonen, M. & Kilpi, T.M. Nasopharyngeal carriage of *Streptococcus pneumoniae* and pneumococcal urine antigen test in healthy elderly subjects. *Scandinavian Journal of Infectious Diseases* **44**, 433-438 (2012).
41. Cardozo, D.M., Nascimento-Carvalho, C.M., Andrade, A.-L.S.S., Silvany-Neto, A.M., Daltro, C.H.C., Brandão, M.-A.S., Brandão, A.P. & Brandileone, M.-C.C. Prevalence and risk factors for nasopharyngeal carriage of *Streptococcus pneumoniae* among adolescents. *Journal of Medical Microbiology* **57**, 185-189 (2008).
42. Tapiainen, T., Paalanne, N., Arkkola, T., Renko, M., Pokka, T., Kaijalainen, T. & Uhari, M. Diet as a risk factor for pneumococcal carriage and otitis media: a cross-sectional study among children in day care centers. *PLoS One* **9**, e90585 (2014).
43. Watson, K., Carville, K., Bowman, J., Jacoby, P., Riley, T.V., Leach, A.J., Lehmann, D. & Team, f.t.K.O.M.R.P. Upper Respiratory Tract Bacterial Carriage in Aboriginal and Non-Aboriginal Children in a Semi-arid Area of Western Australia. *The Pediatric Infectious Disease Journal* **25**, 782-790 (2006).
44. Millar, E.V., O'Brien, K.L., Zell, E.R., Bronsdon, M.A., Reid, R. & Santosham, M. Nasopharyngeal Carriage of *Streptococcus pneumoniae* in Navajo and White Mountain Apache Children Before the Introduction of Pneumococcal Conjugate Vaccine. *The Pediatric Infectious Disease Journal* **28**, 711-716 (2009).

45. Lynch, J.P.I. & Zhanel, G.G. *Streptococcus pneumoniae*: epidemiology and risk factors, evolution of antimicrobial resistance, and impact of vaccines. *Current Opinion in Pulmonary Medicine* **16,** 217-225 (2010).
46. Azzari, C., Moriondo, M., Di Pietro, P., Di Bari, C., Resti, M., Mannelli, F., Esposito, S., Castelli-Gattinara, G., Campa, A., de Benedictis, F.M., Bona, G., Comarella, L., Holl, K. & Marchetti, F. The burden of bacteremia and invasive diseases in children aged less than five years with fever in Italy. *Ital J Pediatr* **41,** 92 (2015).
47. Davidson, M., Parkinson, A.J., Bulkow, L.R., Fitzgerald, M.A., Peters, H. & Parks, D.J. The Epidemiology of Invasive Pneumococcal Disease in Alaska, 1986-1990 Ethnic Differences and Opportunities for Prevention. *Journal of Infectious Diseases* **170,** 368-376 (1994).
48. Ortqvist, A., Hedlund, J. & Kalin, M. *Streptococcus pneumoniae*: epidemiology, risk factors, and clinical features. *Semin Respir Crit Care Med* **26,** 563-74 (2005).
49. Madhi, S.A., Adrian, P., Kuwanda, L., Cutland, C., Albrich, W.C. & Klugman, K.P. Long-Term Effect of Pneumococcal Conjugate Vaccine on Nasopharyngeal Colonization by *Streptococcus pneumoniae*—and Associated Interactions with *Staphylococcus aureus* and *Haemophilus influenzae* Colonization—in HIV-Infected and HIV-Uninfected Children. *Journal of Infectious Diseases* **196,** 1662-1666 (2007).
50. van der Poll, T. & Opal, S.M. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *Lancet* **374,** 1543-56 (2009).
51. Fine, M.J., Smith, M.A., Carson, C.A. & et al. Prognosis and outcomes of patients with community-acquired pneumonia: A meta-analysis. *JAMA* **275,** 134-141 (1996).
52. Hoban, D.J., Doern, G.V., Fluit, A.C., Roussel-Delvallez, M. & Jones, R.N. Worldwide prevalence of antimicrobial resistance in *Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis* in the SENTRY Antimicrobial Surveillance Program, 1997-1999. *Clin Infect Dis* **32 Suppl 2,** S81-93 (2001).
53. Morens, D.M., Taubenberger, J.K. & Fauci, A.S. Predominant Role of Bacterial Pneumonia as a Cause of Death in Pandemic Influenza: Implications for Pandemic Influenza Preparedness. *The Journal of infectious diseases* **198,** 962-970 (2008).
54. Marriott, H.M., Hellewell, P.G., Cross, S.S., Ince, P.G., Whyte, M.K.B. & Dockrell, D.H. Decreased Alveolar Macrophage Apoptosis Is Associated with Increased Pulmonary Inflammation in a Murine Model of Pneumococcal Pneumonia. *The Journal of Immunology* **177,** 6480-6488 (2006).
55. Hirst, R.A., Kadioglu, A., O'Callaghan, C. & Andrew, P.W. The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clinical & Experimental Immunology* **138,** 195-201 (2004).
56. Bogaert, D., Veenhoven, R.H., Sluijter, M., Wannet, W.J., Rijkers, G.T., Mitchell, T.J., Clarke, S.C., Goessens, W.H., Schilder, A.G., Sanders, E.A., de Groot, R. & Hermans, P.W. Molecular epidemiology of pneumococcal colonization in response to pneumococcal conjugate vaccination in children with recurrent acute otitis media. *J Clin Microbiol* **43,** 74-83 (2005).
57. Lindstrand, A., Bennet, R., Galanis, I., Blennow, M., Ask, L.S., Dennison, S.H., Rinder, M.R., Eriksson, M., Henriques-Normark, B., Ortqvist, A. & Alfven, T. Sinusitis and pneumonia hospitalization after introduction of pneumococcal conjugate vaccine. *Pediatrics* **134,** e1528-36 (2014).
58. Capdevila, O., Pallares, R., Grau, I. & et al. Pneumococcal peritonitis in adult patients: Report of 64 cases with special reference to emergence of antibiotic resistance. *Archives of Internal Medicine* **161,** 1742-1748 (2001).

59. Raad, J. & Peacock Jr, J.E. Septic arthritis in the adult caused by *Streptococcus pneumoniae*: A report of 4 cases and review of the literature. *Seminars in Arthritis and Rheumatism* **34,** 559-569 (2004).
60. Watson, D.A. & Musher, D.M. Interruption of capsule production in *Streptococcus pneumonia* serotype 3 by insertion of transposon Tn916. *Infect Immun* **58,** 3135-8 (1990).
61. Blomberg, C., Dagerhamn, J., Dahlberg, S., Browall, S., Fernebro, J., Albiger, B., Morfeldt, E., Normark, S. & Henriques-Normark, B. Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. *J Infect Dis* **199,** 1032-42 (2009).
62. Cremers, A.J., Kokmeijer, I., Groh, L., de Jonge, M.I. & Ferwerda, G. The role of ZmpC in the clinical manifestation of invasive pneumococcal disease. *Int J Med Microbiol* **304,** 984-9 (2014).
63. Browall, S., Norman, M., Tångrot, J., Galanis, I., Sjöström, K., Dagerhamn, J., Hellberg, C., Pathak, A., Spadafina, T., Sandgren, A., Bättig, P., Franzén, O., Andersson, B., Örtqvist, Å., Normark, S. & Henriques-Normark, B. Intraclonal Variations Among *Streptococcus pneumoniae* Isolates Influence the Likelihood of Invasive Disease in Children. *The Journal of Infectious Diseases* **209,** 377-388 (2014).
64. Klugman, K.P., Bentley, S.D. & McGee, L. Determinants of Invasiveness Beneath the Capsule of the Pneumococcus. *Journal of Infectious Diseases* **209,** 321-322 (2014).
65. Pletz, M.W., Maus, U., Krug, N., Welte, T. & Lode, H. Pneumococcal vaccines: mechanism of action, impact on epidemiology and adaption of the species. *International Journal of Antimicrobial Agents* **32,** 199-206 (2008).
66. Hanage, W.P., Finkelstein, J.A., Huang, S.S., Pelton, S.I., Stevenson, A.E., Kleinman, K., Hinrichsen, V.L. & Fraser, C. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics* **2,** 80-84 (2010).
67. Reinert, R.R. The antimicrobial resistance profile of *Streptococcus pneumoniae*. *Clinical Microbiology and Infection* **15,** 7-11 (2009).
68. Slavin, R.G., Spector, S.L., Bernstein, I.L., Slavin, R.G., Kaliner, M.A., Kennedy, D.W., Virant, F.S., Wald, E.R., Khan, D.A., Blessing-Moore, J., Lang, D.M., Nicklas, R.A., Oppenheimer, J.J., Portnoy, J.M., Schuller, D.E., Tilles, S.A., Borish, L., Nathan, R.A., Smart, B.A. & Vandewalker, M.L. The diagnosis and management of sinusitis: A practice parameter update. *Journal of Allergy and Clinical Immunology* **116,** S13-S47 (2005).
69. Cober, M.P. & Phelps, S.J. Penicillin Prophylaxis in Children with Sickle Cell Disease. *The Journal of Pediatric Pharmacology and Therapeutics : JPPT* **15,** 152-159 (2010).
70. Mandell, L.A., Wunderink, R.G., Anzueto, A., Bartlett, J.G., Campbell, G.D., Dean, N.C., Dowell, S.F., File, T.M., Musher, D.M., Niederman, M.S., Torres, A. & Whitney, C.G. Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults. *Clinical Infectious Diseases* **44,** S27-S72 (2007).
71. Hakenbeck, R., Brückner, R., Denapaite, D. & Maurer, P. Molecular mechanisms of β-lactam resistance in *Streptococcus pneumoniae*. *Future Microbiology* **7,** 395-410 (2012).
72. Griffith, F. The Significance of Pneumococcal Types. *The Journal of Hygiene* **27,** 113-159 (1928).
73. Downie, A.W. Pneumococcal Transformation-A Backward View Fourth Griffith Memorial Lecture. *Microbiology* **73,** 1-11 (1972).
74. Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson,

W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C., Dougherty, B.A., Morrison, D.A., Hollingshead, S.K. & Fraser, C.M. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293,** 498-506 (2001).

75. Croucher, N.J., Coupland, P.G., Stevenson, A.E., Callendrello, A., Bentley, S.D. & Hanage, W.P. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* **5,** 5471 (2014).

76. Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E.I. & Orihuela, C.J. Identification of a Candidate *Streptococcus pneumoniae* Core Genome and Regions of Diversity Correlated with Invasive Pneumococcal Disease. *Infection and Immunity* **74,** 4766-4777 (2006).

77. Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P. & Bentley, S.D. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46,** 305-9 (2014).

78. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* **11,** 435-42 (2013).

79. Duffield, M., Cooper, I., McAlister, E., Bayliss, M., Ford, D. & Oyston, P. Predicting conserved essential genes in bacteria: *in silico* identification of putative drug targets. *Mol Biosyst* **6,** 2482-9 (2010).

80. Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E., Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., Turner, P., Bentley, S.D. & Parkhill, J. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet* **10,** e1004547 (2014).

Chapter 2

Advances and perspectives in computational prediction of microbial gene essentiality

Fredrick M. Mobegi
Aldert Zomer
Marien I. de Jonge
Sacha A. F. T. van Hijum

Briefings in Functional Genomics 2016; doi: 10.1093/bfgp/elv063.

Abstract

The minimal subset of genes required for cellular growth, survival, and viability of an organism are classified as essential genes. Knowledge of essential genes gives insight into the core structure and functioning of a cell. This might lead to more efficient antimicrobial drug discovery, to elucidation of the correlations between genotype and phenotype, and a better understanding of the minimal requirements for a (synthetic) cell. Traditionally, constructing a catalog of essential genes for a given microbe involved costly and time-consuming laboratory experiments. While experimental methods have produced abundant gene essentiality data for model organisms like *Escherichia coli* and *Bacillus subtilis*, the knowledge generated cannot automatically be extrapolated to predict essential genes in all bacteria. In addition, essential genes identified in the laboratory are by definition 'conditionally essential', as they are essential under the specified experimental conditions: these might not resemble conditions in the microorganisms' natural habitat(s). Also, large-scale experimental assaying for essential genes is not always feasible due to the time investment required to setup these assays. The ability to rapidly and precisely identify essential genes *in silico* is therefore important and has great potential for applications in medicine, biotechnology, and basic biological research. Here, we review the advances made in the use of computational methods to predict microbial gene essentiality, perspectives for the future of these techniques, and the possible practical applications of essential genes.

Introduction

Inactivation of essential genes in an otherwise wild type organism results in lethality. These genes therefore represent the foundation of cellular life [1, 2]. Identifying essential genes is therefore valuable and important in biology, industrial bioprocessing, and medicine. For example, it could aid in comprehending the basic principles behind how cells function [3], and the complex relations between genotype and phenotype [4], which are fundamental questions in biology and genetics. Understanding the function of essential genes is prerequisite to discovering the core components of a minimal cell [5], potentially facilitating re-engineering of microorganisms [6] with desired phenotypical traits for research and biotechnology. Additionally, since essential genes confer lethal phenotypes to microorganisms when deleted or inactivated, they form promising drug targets upon which potent antibiotics could be developed [7, 8]. Knowledge of gene essentiality has been applied in discovering candidate 'human disease genes' (genes with disease associated alleles), their mode of inheritance, and contribution to developmental abnormalities or disease [9].

Essential genes have been identified in a number of model organisms [10-15]. However as recently reviewed [2], several studies querying gene essentiality in the same organisms under similar experimental conditions have produced different catalogs of essential genes. This lack of consensus makes it challenging to determine gene essentiality in model organisms, let alone in non-model or poorly researched organisms. The differences are possibly a result of 'conditional or contextual' essentiality: the essentiality of a gene depends on its context, which might be a defined growth media or conditions, genetic context, or a particular developmental stage of a microorganism [16]. Moreover, the longer timespans required for conducting experiments also give enough time for isozymes to be upregulated significantly affecting essentiality prediction. Most studies have consistently deciphered essential genes under rich media conditions (Supplementary Table 1); in other words, in the richness of a full complement of vital nutrients and devoid of environmental stress [8, 11, 13, 14, 17]. Although laboratory rich media conditions are undoubtedly not a proxy of conditions in a microorganism's natural niche, essential genes determined under these conditions provide a near-complete representation of genes needed in most *in situ* niches [11]. Therefore, for the purpose of this review, we define the "essentiality" of a gene as its indispensability under rich media conditions.

Gene essentiality studies have advanced significantly in the past few years owing to a plethora of *in vitro*, *in vivo* (laboratory), and *in silico* methods. Laboratory methods assess gene essentiality by observing lethal phenotypes ensuing from random or systematic gene inactivation using transposon mutagenesis [12], gene knockouts [11, 18], genetic complementation [19], and RNA interference [20]. However, genomic-scale discovery of essential genes using laboratory techniques is often complex, costly, time consuming,

and is contextual since it can be influenced by growth conditions as well as genetic context [16]. Therefore, to establish accurate results, a consensus of predicted essential genes across multiple laboratories is required. To circumvent these complexities, *in silico* techniques have been developed to predict essential genes [21-23]. Computational methods have gained popularity over the past years for numerous reasons. First, computational methods are less time consuming and they benefit from knowledge obtained from other organisms. The essential genes identified from several microorganisms provide seed information for training gene essentiality predictors for less researched organisms. Second, the abundance of 'omics' data from genomic sequencing projects provides opportunities for microbial functional genomics. Lastly, bioinformatics has greatly developed over recent years, significantly advancing tools available to discover essential genes in sequenced genomes. It is noteworthy that computational methods cannot (yet) predict conditional essentiality but rather predict whether a gene is essential or not.

In this review, we focus on advances made in genome-wide microbial gene essentiality prediction, particularly using computational methods. We discuss the fundamental principles of computational gene essentiality prediction tools, and provide an opinion on the choice of method. We also explore the possible practical applications of essential genes and give a perspective into the future of computational methods in predicting gene essentiality.

Computational techniques for gene essentiality prediction

Many *in silico* prediction methods have been established to aid in *post hoc* analysis of experimental readouts, or mining 'omics' data for encoded signatures to identify essential genes. Below we discuss approaches commonly used to predict gene essentiality (Figure 1). They commonly analyze intrinsic genomic features, such as localization signals, codon adaptation indices, GC content, gene orthologs, rate of gene evolution, and phyletic gene retention [21, 22, 24]. Other integrated approaches such as network analysis [3, 25] and machine learning on combinations of features and approaches [24, 26] are also discussed.

Transposon sequencing methods

Transposons have been widely used in techniques like signature tagged mutagenesis [27] to manipulate genes in various microorganisms [12, 28, 29] albeit with low resolution. Recently however, various high-throughput techniques including Tn-seq [15], INSeq [30], HITS [31], TraDIS [32], and variants thereof have harnessed the power of traditional transposon mutagenesis, next-generation sequencing (NGS), and post-hoc *in silico* tracking of the insertions, to explore gene function and higher-order genome organization [14].

Transposon sequencing and analysis (TSA) techniques commonly rely on the construction of transposon mutant libraries in which non-essential genes contain transposon insertions, followed by growth of the mutant libraries in defined *in vitro* or *in vivo* (e.g. host infection models) conditions. The relative frequency of each mutant in the population at the beginning and the end of the experiment is then determined by means of NGS at the transposon junctions. Genes that are essential for growth under a particular condition will not accumulate transposon insertions. From this data, the fitness of every gene to the experimental conditions to which the transposon libraries were subjected is quantified [15, 30-32]. The relatedness and differences between various TSA techniques have comprehensively been reviewed [33, 34]. Their main advantages are the high levels of accuracy and sensitivity in predicting gene essentiality, and their ability to be adapted for analyses in a wide range of species. By using certain regimes to store mutant libraries, it is also possible to obtain strains with desired gene knockout(s). In addition, the sequencing protocols used generate short sequence reads of millions of DNA molecules simultaneously, allowing whole genomes to be investigated in a single experiment. Nonetheless, TSA techniques are dependent on strong molecular amenability of an organism to allow creation of saturated mutant libraries and accurate deep sequencing [8], making them quite expensive for routine use. For this reason, computational approaches like homology mapping and machine learning, which may rely solely on computer-mined essentiality determinants would be desirable (Figure 1).

Homology mapping models

The term 'homologs' refers to two or more genes related by descent from a common ancestral DNA sequence. This relationship may arise between genes separated by speciation (orthologs) or genetic duplication (paralogs). Prediction of essential genes based on sequence homology, especially to known essential genes, is arguably the simplest and earliest used method in the genomic era. Shortly after the availability of the first two completely sequenced bacterial genomes, homology models were employed to predict gene essentiality [35], and to establish the minimal genome [17] in *Haemophilus influenzae* and *Mycoplasma genitalium*. These models rely on heuristic algorithms embedded in sequence alignment programs like Muscle [36], Clustal [37-39], T-Coffee [40], and database search tools like BLAST, to compare query sequences with a library or database of subject sequences whose essentiality is known. Sequences that are similar above defined percentage identity and *e-value* threshold, and length coverage are grouped as homologs. Homology models show high confidence levels owing to these metric thresholds.

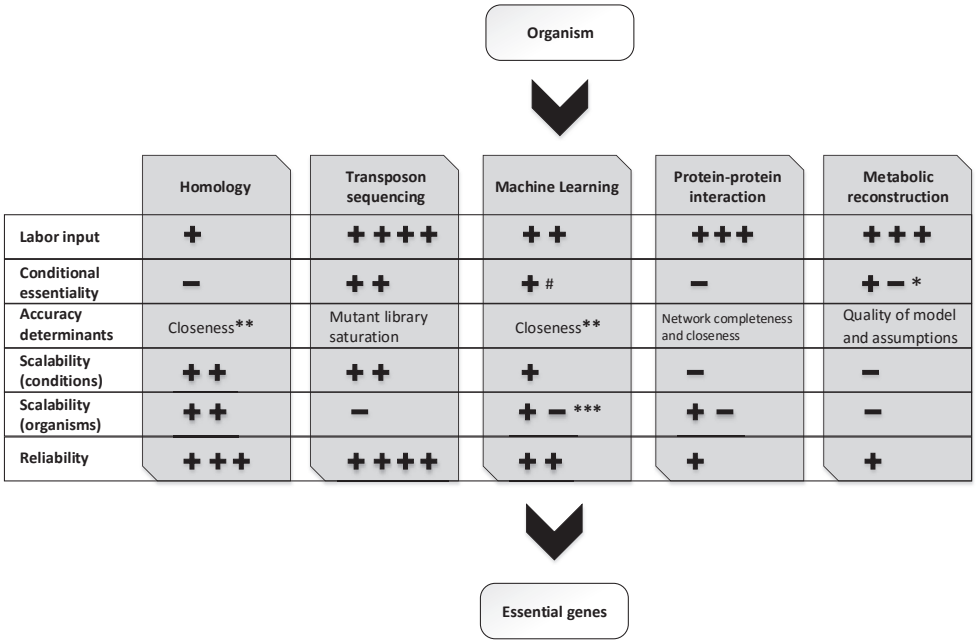


Figure 1: Summary of the computational methods used in predicting essential genes.
#The ability to identify conditionally essential genes is usually specific for the training dataset.
*****Prediction of conditional essentiality is influenced by the quality of the input model, objective function, among other factors.
**In homology models, the query and subject sequences should have maintained enough closeness throughout evolution. In ML models, the study subjects should be close enough (e.g. same or close species) to the training set, hence disadvantageous while dealing with distantly related species.

Essential genes evolve slowly and tend to be more conserved than non-essential genes in bacteria [41]. Selection on essential genes is more stringent than on non-essential genes, increasing the average likelihood that orthologues of essential genes are conserved in bacteria [1, 42] and almost certainly essential. This allows extrapolation of essentiality from one member to an entire group of homologous genes. Many bacterial genomes are publicly available from genome sequencing projects (Figure 2). Various homology prediction tools and databases that collate homologous proteins (Supplementary Table 2) are also publicly available. This has greatly simplified determining genes that share ancestry, making homology models attractive for predicting gene essentiality based solely on genomic sequences. However, they have various limitations. First, they are limited to conserved orthologs between species, which often account for a small portion of the genome [43]. Moreover, since the model only considers computationally determined orthologous genes based on sequence similarity, highly evolving genes may

be overlooked, consequently leading to underestimation of essential genes in a genome [5]. Secondly, orthologues, especially in distantly related species, often show variations in gene regulations, post-translational protein modification, divergence in cellular pathways, redundancies in processes, gene duplications, and other niche specializations [44], leading to potential multiplicity in relative gene essentiality. For example, Hutchison and colleagues [28] successfully used transposon insertions to disrupt some of the 256 essential genes predicted using homology [35], suggesting that they are possibly non-essential. Yu *et al.,* also identified 787 non-essential *S. sanguinis* genes, which had orthologs in all 48 *Streptococcus* genomes they analyzed [45]. Additionally, the gene encoding alanyl-tRNA synthetase (*alaS*) is essential in *Escherichia coli* but not in *Pseudomonas aeruginosa*: probably due to functional redundancy caused by a paralog, PA2106, in *Pseudomonas* [26]. For this reason, absence of paralogs is in generally thought to be a strong indicator of essentiality in cross-species homology analyses [8, 46]. Although it is more straight-forward to predict essentiality for single copy genes, paralogous genes could still be essential: deletion of all copies of a duplicated gene that encodes an essential function should lead to lethality. Finally, notwithstanding the tendency of essential genes to be highly conserved, genes conserved across species are not always essential. Indeed, in model organisms, only less than 25% of all conserved genes have experimentally been validated to be essential [29, 47], indicating that similarity of sequences does not always warrant extrapolation of the essentiality annotation amongst homologs.

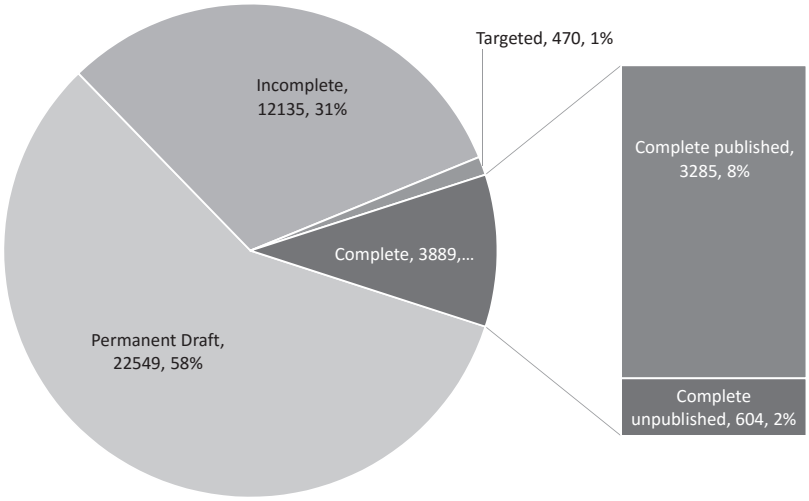


Figure 2: The genome project coverage for bacteria**.** Approximately 39,442 bacterial genome projects are documented according to GOLD (Genome OnLine Database; http://www.genomesonline.org) as of June 30, 2015.

Protein network topology models

Two or more proteins can establish physical contact (protein-protein interactions; PPIs) as a result of biochemical events and/or electrostatic forces. These interactions are mainly determined from quantum chemistry, molecular dynamics, signal transduction, and biochemistry assays among others [48-51] and in certain cases can be used to predict essentiality. The abundance of experimental data generated from small-scale analyses and high-throughput procedures have assisted in defining PPIs within the interactome (all possible molecular interactions within a cell). Large-scale exploration of the topological properties of these networks is important in understanding the organizational and functional principles of individual proteins in biological pathways [25], and consequently their essentiality. Various descriptors for centrality of a node in a network have been effectively applied to identify essential proteins in PPIs [25, 52, 53]. For example, deletion of a hub (highly linked) protein is more likely to lethally perturb the network than deletion of a non-hub (peripheral) protein [54]. It is therefore widely agreed that hub proteins evolve slowly and are most likely to be essential [53]. However, prediction of gene essentiality in less studied genomes using protein networks is expensive and arduous, primarily due to the limitations of experimental data (containing missing values, false positives, false negatives, or differing between replicates - suggesting that the data is erroneous, incomplete, or both) necessary to build and characterize PPIs [55], and the complexities in computational inference of PPI networks [56]. Their accuracy also depend on the completeness of the network and they cannot be used to predict conditional essentiality. Moreover, proteins that lack known interactions with other proteins are completely disregarded in PPI models.

Metabolic network reconstruction and simulation models

While studying PPI networks gives a basic understanding of gene and protein interactions, they are limited in elucidating the complex and dynamic interactions among molecular components of cellular networks at genome-scale. Whole genome metabolic network reconstruction under constraint-based reconstruction and analysis (COBRA) scaffold [57, 58] allows for an in-depth insight into the metabolic capabilities of an organism, particularly in correlating the genome with molecular physiology [59]. These methods are based on the following fundamental concepts: (1) the burden of physicochemical constraints to limit quantifiable phenotypes, (2) identification and algebraic account of evolutionary selective pressures, and (3) a genome-scale perception of cell metabolism that accounts for all cellular metabolic gene products [60]. Metabolic networks for many microorganisms have been reconstructed [61] or can be reconstructed and, to some extent, curated using automated systems such as model SEED [62], RAST [63], and BiGG [64]. Also, by integrating homology modeling, genome-scale models that show substantial predictive power in auxotrophy and essentiality predictions have been created for multiple strains of lesser studied organisms by starting with the genome-scale model of a well-studied organism like *Escherichia coli* K12 MG1655 [65]. Nonetheless,

significant efforts are required to manually curate and ascertain the reliability of automatically generated metabolic models to improve their reliability for gene essentiality prediction. It is noteworthy that following the immense success in metabolic networks reconstruction, significant efforts are being made to model transcriptional networks [66], signaling networks [67], and computing of protein expressions needed to perform metabolism and proteome synthesis [68]. These efforts are providing crucial input to extend gene essentiality prediction using network reconstruction modelling from metabolism to incorporate other non-metabolic cellular processes.

Flux balance analysis (FBA) is the most commonly used constrained-based approach for *in silico* prediction of microbial phenotypes from metabolic models [69, 70]. It integrates biochemical constraints with the stoichiometry of metabolic and transport reactions, their reversibility and subcellular localization, thereby reducing the intricacy of potential dynamic states in order to predict metabolite fluxes at steady state. FBA has been to simulate gene knock-out and evaluate the associated lethality on the system, enabling the identification of essential genes [70]. For a given objective function and each *in silico* gene deletion, essentiality is evaluated by calculating the optimal production of defined biosynthetic precursors: identified auxotrophic requirements and impaired functions (indicating simulated gene knockouts which inhibit *in silico* production of precursors contained within the objective function e.g. alanine production) are classified as essential for the objective function. FBA relies on stoichiometric characteristics and does not require kinetic parameters (which are often difficult to obtain) allowing it to be used on any fully sequenced and annotated organism [69]. Nevertheless, FBA has important limitations: first, while FBA could be integrated with modal analyses at steady-state, it cannot be used to investigate genome-scale metabolic reactions under transient dynamic states without including data on enzyme kinetics [71]. Secondly, FBA cannot be used to directly predict immediate suboptimal flux states and metabolite concentration following a genetic perturbation. Organisms naturally adapt to perturbations by readjusting various regulatory mechanisms, enzyme expressions, and fluxes to bypass the effects. Such immediate changes and the effect of regulatory mechanisms cannot be explicitly specified in FBA. Some of these limitations have been addressed in variants of FBA such as MOMA [72], ROOM [73], MEA [74] and dynamic FBA [71]. Lastly, FBA sometimes disagrees with experimental data; these discrepancies could be addressed by the addition of enzyme reactions through "gap filling" [69, 75]. Overall, given the substantial input required and the inability to provide direct readout for conditionally essential genes, metabolic network reconstruction models are undesirable first-choice methods for exploring gene essentiality in novel genomes.

Integrated features machine learning models

Integrative machine learning (ML) models rely on constructing and training a classifier for predicting gene essentiality. They integrate multiple characteristics or features encoded in an organism's genomic sequence, which are known to be associated with essentiality

[26, 76]. The classifiers are trained and tested using well-annotated genomes, then applied to identify putatively essential genes in other (novel) genomes [23-25]. The ever-increasing number of experimentally determined essential genes has improved the understanding of distinguishing properties of essential genes. As a result, it is possible to easily select features towards improving the predictive accuracy of machine learning models, making them less laborious. Predictive accuracy of ML classifiers resulting from a combination of different features may vary but no specific combinations have been confirmed to be optimally robust. The reliability of ML models however depend on the closeness of the training dataset to the study dataset. Normally, ML models may be prone to overfitting, potentially allowing irrelevant information or noise to be presented as valid predictions. Domingos and colleagues reviewed overfitting as well as other sources of errors in ML, and the possible methods of combating them [77]. These models may not be suitable for predicting conditional gene essentiality. Various experimental, genomic and protein features have been used to train and build classifier for genes essentiality prediction in different studies (Table 1). However, no single study has reported use of all the features in a single predictive model to predict gene essentiality. The features are often used selectively based on their accuracy and whether they can patently be determined for the organism under study.

Applications of essential genes

Discovering potential drug targets

Several diseases are becoming increasingly difficult to control due to the emergence of drug-resistant pathogenic strains, necessitating a search for new antimicrobials. Identification and prioritization of drug targets in novel pathogens is the initial and one of the most important steps during drug discovery (Figure 3). Understanding the functions of the target proteins, and consequently the mechanisms of action (MOA) are important to design putative inhibitors. As such, drug target discovery sets a foundation for developing drugs with desired therapeutic properties. Inhibiting essential proteins will confer bacteriostatic or bactericidal effects. They therefore form promising targets for discovering potent antibiotics against novel pathogens [7].

Table 1: Features that can be used for *in silico* prediction of gene essentiality

Feature	Rationale	Reference
Gene expression profile [a]	Genes that are not expressed under given conditions are less likely to be essential Co-expressed genes are often involved in the same pathway or similar cellular function Interacting proteins are frequently co-expressed	[95]
Protein localization and biological processes (enrichment of Gene Ontology (GO) [b]	Essential proteins are enriched in, but not exclusive to the cytoplasm: compared to essential genes, significantly higher proportions of non-essential genes are located in the cytoplasmic membrane, periplasm, outer membrane, cell wall, and extracellularly Gene Ontology (GO) term transcriptional regulation annotations is enriched in essential genes	[21, 96]
Functional domains	The functional units of proteins are domains, most of which are highly conserved in diverse genera	[97, 98]
Total upstream gene size [c]	Genes with larger upstream sizes (promoter regions) are significantly underrepresented in indispensable genes Genes regulated by multiple transcriptional regulators are likely to have larger upstream regions in order to house the various cis-regulatory elements Genes with more complex regulation are generally dispensable	[3, 21, 42, 99]
Phyletic retention measure [d]	Essentiality of a gene is extrapolated if the annotated function of that gene can be detected in different genera as opposed to sequence similarity Specificity increases with inclusion of diverse genera in the analysis	[21, 45]
GC content	Commonly used to identify genes that are essential under high temperature selection. The DNA double helix is stabilized primarily by hydrogen bonds between nucleotides and base-stacking interactions among aromatic nucleobases: the GC pair contains three hydrogen bonds, whereas the AT pairs contain two. DNA with high GC-content is believed to be more robust and stable.	[100]
Codon usage	The probability of a deleterious substitution in essential proteins is expected to be negligible, resulting in lower nonsynonymous substitution rates	[41]
Orthology and paralogy	In bacteria, essential genes are generally more conserved across species (orthologs) than non-essential genes Duplicated genes within a genome (paralogues) are also less likely to be essential because the duplicate gene serves as a backup and can replace the original copy. Inactivation of both copies may however result to lethality.	[41, 46]
Protein connectivity	Highly connected proteins in a network evolve slowly and are more likely to be essential- see PPI networks	[53, 54]
Strand bias	Essential genes tend to be encoded on the leading strand of the circular chromosome	[101]

[a] An organism's genetic code is interpreted by gene expression into functional gene products: Properties of gene expression give rise to a phenotype, often expressed by synthesis of proteins that act as catalytic enzymes in specific metabolic pathways, or control the organism's physical traits [102].
[b] Essential functional proteins domains have also been identified and used to predict gene essentiality [97, 98].
[c] The connection between regulation complexity, intergenic distance, and gene essentiality has been shown in *Drosophila melanogaster* and *Caenorrhabditis elegans* [99]. Since transcription factor binding sites in the promoter region are discovered using laborious experimental methods, the possibility of using easy-to-determine upstream region size, as a representation for regulatory complexity in integrative models, is advantageous.
[d] Often confused with conservation; a measure of substitution rate, phyletic gene retention is a measure of the number of organism in which an ortholog is present [21, 45]. It is therefore assumed that most essential genes could be predicted based on the genome annotations. However, the number of essential genes is likely to decrease with increased diversity in the genera, subsequently leading to under-prediction [45].

In our recent study, using a high-throughput genome wide screening approach, we identified essential genes in bacterial respiratory pathogens [8]. From these essential genes, additional criteria were applied to prioritize, and experimentally validate some potential target proteins and pathways that can be modulated by bioactive agents. Despite the intensifying research efforts, adoption of the biomedical discoveries into developmental stages of drug discovery, and subsequently into marketable products has been dismal. Indeed, more than 80% of all potential products going into the drug development pipeline never make it to the market [78]. The problem might partly be due to the lack of comprehensive biochemical knowledge of the drug targets and the MOA of their "lead compound" inhibitors; such that, unexpected biological effects are not fully assessed prior to clinical trials [79]. Moreover, taking a drug through research and development to clinical approval requires immense investments in both time and cost further exacerbating the problem. In fact, only a handful of therapeutic molecules have been approved in recent years by the regulatory agencies in the United States and Europe [78].

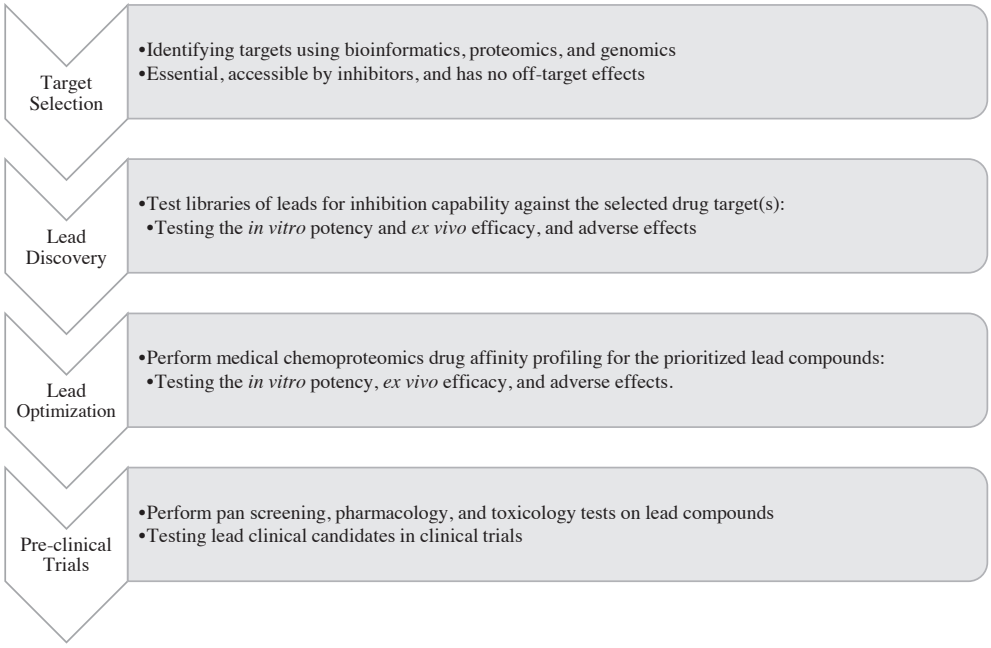


Figure 3: Schematic representation of the drug discovery pipeline.

Food microbiology and industrial bioprocessing

Numerous food products, including ripened cheese, pickles, wine, beer, bread, yoghurt and other fermented foods, owe their production and characteristics to microorganisms [80]. These foods are to some extent naturally preserved due to the fermentation process. Concurrently, their shelf-life is prolonged significantly over that of the raw materials from which they are manufactured. Several biopolymers produced by microorganisms are also used in the food industry [81]. In addition, "generally recognized as safe; GRAS" bacteria or probiotics are becoming increasingly vital in the food industry [82]. However, microorganisms also constitute potential food and process contaminants, and foodborne pathogens [83]. Understanding essential genes is therefore invaluable in optimizing various facets of industrial fermentation and bioprocessing.

First, given a growth medium containing a defined carbon source (raw product), metabolic reconstruction can be used to predict the best processing conditions required for a given microorganism to maximize production. The ability to evaluate interactions between microorganisms and relevant metabolic capabilities in new microbes also provides leads for novel and safe starter cultures. Moreover, the knowledge gives an insight into all (conditionally) essential metabolic pathways involved in producing a given product and co-expressed nonessential pathways which, if feasible, could be inactivated to improve efficiency. For example, attempts have been made in using microbes to produce alginate, a product conventionally isolated from farmed brown seaweed [84]. Knowledge of (conditional) gene essentiality was used to pinpoint the interactions between multiple microorganisms during the course of the fermentation. To facilitate product optimization strategies, catalogs of (conditionally) essential genes for specific microorganisms in specific food products can aid in establishing biobanks and biorepositories. The biobanks can be screened for new "safe" microbes with desired phenotypic traits and predicted interactions for a given fermentation, or to create novel fermented products.

Secondly, while most foodborne pathogens or food spoilage bacteria and industrial contaminants are cleared by standard sterilization, cooking and preservatives, including bacterial toxins like nisin, studies have shown that some potentially harmful microorganisms can survive these conventional food processing methods [85, 86]. Food samples could be tested by PCR for the presence of general spoiler marker genes indicating a possible contamination. Such tracking tests are significantly faster than conventional techniques. Moreover, unique and essential spoiler genes could be considered as prime targets for bespoke decontamination strategies without inversely altering the production pipeline.

Lastly, genomics-based determination of gene essentiality also generates valuable knowledge that can be used for metabolic engineering, optimizing cell factories and development of novel preservation methods, provided that these solutions are ethically acceptable.

Bioremediation

Compared to conventional physicochemical strategies, microbes provide a safe and cheap alternative for environmental remediation, pollution prevention, and waste treatment [87]. Although highly diverse and specialized microbial populations present in the environment efficiently eliminate many pollutants, the process is normally quite slow, potentially permitting pollutants to accumulate above hazardous levels. For example, bioremediation of the *Exxon Valdez* oil spill in Alaska using indigenous microflora was cost effective and scientifically rational [88]. However, fertilizers had to be applied in order to accelerate the process. The fertilizers present a separate environmental imbalance albeit minimal compared to the oil spill. Unlike oil whose constituent hydrocarbons are largely biodegradable, most recalcitrant compounds, especially heavy metals, contain structural elements or substituents that seldom occur in nature. Because of the rarity of these compounds, currently known microorganisms have probably not evolved appropriate pathways to bioaccumulate them. While some xenobiotics are inefficiently or incompletely biotransformed, or their complex mixtures inhibit degradation by existing pathways, for others, derivative pathways have not been described [89]. Knowledge of (conditional) gene essentiality can therefore aid in identifying novel biodegradation pathways in (new) microorganisms. Additionally, the knowledge could facilitate genetic modification of microbes to broaden their substrates range, successfully enhancing cleanup while producing specialized (end- or by-) products with less ecological harm.

Genotype-phenotype correlation

Mendel's classical observations of varied phenotypes in peas conjured a paradigm of distinct alterations in an organism's DNA (genotype) that cause disruptions in gene function and characteristic phenotype. Ever since, phenotypes have been used to systematically discover their plausible genetic background. However, the phenotype of a given strain is not only a product of its gene content but also its cellular regulatory mechanisms [90] and environmental factors [91]. Although genotype-phenotype association studies do not factor in the effects of regulatory mechanisms, they allow for straight-forward screening of candidate genotype to phenotype relationships. Additionally, the natural diversity and adaptive responses of microbial strains to environmental changes could also be investigated using knowledge of conditional gene essentiality. For example, using transcriptomic diversity between strains of *Lactococcus lactis* isolated from diary and nondairy niches, the basis of phenotypic differences observed in fermented food products at the level of acidification properties has been investigated [92].

Perspectives

Gene essentiality prediction using computational methods will become more important with the ongoing advances in biology. Expanding computational methods to predict conditionally essential genes, which are currently predicted exclusively using laboratory techniques, may soon be realized. Evidently, predicting conditional essentiality requires many experimentally determined features that cannot be determined computationally yet. *In silico* reconstruction of microbial genomes with preferred phenotypic traits also stands to benefit. In fact, the *ab initio* assembly of a synthetic cell [6], and genome transplantation [93] have been accomplished in *Mycoplasma mycoides.* Fabricating viable cells that harbor housekeeping functions and only genes encoding desired phenotypes is therefore achievable and can be perfected in the future. There have been genome engineering attempts to improve *de novo* biosynthesis of vanillin [94]. With the global demand for natural food ingredients, flavors, fragrances, biopolymers, and drugs increasing rapidly, specialized fabricated microbes that perform "natural-like" bioconversions more efficiently might be desirable. Such projects will undoubtedly revolutionize processes beyond current technologies when they are scaled up to industrial size production. Additionally, by creating, testing and optimizing specialized genetic circuits, our understanding of cell biology will also advance significantly. In conclusion, it is our believe that future studies could build on the knowledge reviewed here, and expand it to improve accuracy and dependability of *in silico* tools in predicting essential genes.

Key points

- Essential genes are fundamental for cellular growth and viability of an organisms.
- They form attractive drug targets and essential components of a minimum cell for biotechnology and basic biological research.
- Supplementing or complementing traditional laboratory gene essentiality prediction methods with high-throughput computational approaches is gaining interest.
- Currently, transposon insertion sequencing is the most reliable but quite expensive method that combines wet-laboratory and computational tracking to predict gene essentiality.
- Solely computational methods including homology models, machine learning models, metabolic network reconstruction, and protein-protein interaction models, are reliable but largely influenced by the quality of data and evolutionary distance between subjects

.

References

1. Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Débarbouillé, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Mauël, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F.M.L., Sekiguchi, J., Sekowska, A., Séror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U. & Ogasawara, N. Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences* **100**, 4678-4683 (2003).
2. Juhas, M., Eberl, L. & Glass, J.I. Essence of life: essential genes of minimal genomes. *Trends in Cell Biology* **21**, 562-568 (2011).
3. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends in Genetics* **20**, 227-231 (2004).
4. Dowell, R.D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D.A., Rolfe, P.A., Heisler, L.E., Chin, B., Nislow, C., Giaever, G., Phillips, P.C., Fink, G.R., Gifford, D.K. & Boone, C. Genotype to Phenotype: A Complex Problem. *Science* **328**, 469 (2010).
5. Gil, R., Silva, F.J., Peretó, J. & Moya, A. Determination of the Core of a Minimal Bacterial Gene Set. *Microbiology and Molecular Biology Reviews* **68**, 518-537 (2004).
6. Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.-Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., Smith, H.O. & Venter, J.C. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52-56 (2010).
7. Chung, B.K.-S., Dick, T. & Lee, D.-Y. In silico analyses for the discovery of tuberculosis drug targets. *Journal of Antimicrobial Chemotherapy* **68**, 2701-2709 (2013).
8. Mobegi, F.M., van Hijum, S.A., Burghout, P., Bootsma, H.J., de Vries, S.P., van der Gaast-de Jongh, C.E., Simonetti, E., Langereis, J.D., Hermans, P.W., de Jonge, M.I. & Zomer, A. From microbial gene essentiality to novel antimicrobial drug targets. *BMC Genomics* **15**, 958 (2014).
9. Dickerson, J.E., Zhu, A., Robertson, D.L. & Hentges, K.E. Defining the Role of Essential Genes in Human Disease. *PLoS ONE* **6**, e27368 (2011).
10. Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N. & Mekalanos, J.J. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci U S A* **99**, 966-71 (2002).
11. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian,

K.-D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. & Johnston, M. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391 (2002).

12. Salama, N.R., Shepherd, B. & Falkow, S. Global Transposon Mutagenesis and Essential Gene Analysis of Helicobacter pylori. *Journal of Bacteriology* **186**, 7926-7935 (2004).
13. Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O. & Venter, J.C. Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 425-430 (2006).
14. Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H. & Shapiro, L. The essential genome of a bacterium. *Mol Syst Biol* **7**, 528 (2011).
15. van Opijnen, T., Bodi, K.L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6**, 767-72 (2009).
16. D'Elia, M.A., Pereira, M.P. & Brown, E.D. Are essential genes really essential? *Trends in Microbiology* **17**, 433-438 (2009).
17. Mushegian, A. The minimal genome concept. *Current Opinion in Genetics & Development* **9**, 709-714 (1999).
18. Roemer, T., Jiang, B., Davison, J., Ketela, T., Veillette, K., Breton, A., Tandia, F., Linteau, A., Sillaots, S., Marta, C., Martel, N., Veronneau, S., Lemieux, S., Kauffman, S., Becker, J., Storms, R., Boone, C. & Bussey, H. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. *Molecular Microbiology* **50**, 167-181 (2003).
19. Andreadaki, M., Morgan, R.N., Deligianni, E., Kooij, T.W.A., Santos, J.M., Spanos, L., Matuschewski, K., Louis, C., Mair, G.R. & Siden-Kiamos, I. Genetic crosses and complementation reveal essential functions for the Plasmodium stage-specific actin2 in sporogonic development. *Cellular Microbiology* **16**, 751-767 (2014).
20. Cullen, L.M. & Arndt, G.M. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunol Cell Biol* **83**, 217-223 (2005).
21. Gustafson, A., Snitkin, E., Parker, S., DeLisi, C. & Kasif, S. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* **7**, 265 (2006).
22. Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M. & Gerstein, M. Predicting essential genes in fungal genomes. *Genome Research* **16**, 1126-1135 (2006).
23. Chen, Y. & Xu, D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* **21**, 575-581 (2005).
24. Deng, J. An Integrated Machine-Learning Model to Predict Prokaryotic Essential Genes. in *Gene Essentiality*, Vol. 1279 (ed. Lu, L.J.) 137-151 (Springer New York, 2015).
25. Plaimas, K., Eils, R. & Konig, R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Systems Biology* **4**, 56 (2010).

26. Deng, J.Y., Deng, L., Su, S.C., Zhang, M.L., Lin, X.D., Wei, L., Minai, A.A., Hassett, D.J. & Lu, L.J. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Research* **39,** 795-807 (2011).
27. Hensel, M., Shea, J., Gleeson, C., Jones, M., Dalton, E. & Holden, D. Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269,** 400-403 (1995).
28. Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O. & Craig Venter, J. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science* **286,** 2165-2169 (1999).
29. Song, J.H., Ko, K.S., Lee, J.Y., Baek, J.Y., Oh, W.S., Yoon, H.S., Jeong, J.Y. & Chun, J. Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol Cells* **19,** 365-74 (2005).
30. Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R. & Gordon, J.I. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6,** 279-89 (2009).
31. Gawronski, J.D., Wong, S.M., Giannoukos, G., Ward, D.V. & Akerley, B.J. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A* **106,** 16422-7 (2009).
32. Langridge, G.C., Phan, M.D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., Wain, J., Parkhill, J. & Turner, A.K. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* **19,** 2308-16 (2009).
33. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* **11,** 435-42 (2013).
34. Barquist, L., Boinett, C.J. & Cain, A.K. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biology* **10,** 1161-1169 (2013).
35. Mushegian, A.R. & Koonin, E.V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences* **93,** 10268-10273 (1996).
36. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32,** 1792-7 (2004).
37. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. *Nucleic Acids Research* **25,** 4876-4882 (1997).
38. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673-4680 (1994).
39. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947-2948 (2007).
40. Notredame, C., Higgins, D.G. & Heringa, J. T-coffee: a novel method for fast and accurate multiple sequence alignment1. *Journal of Molecular Biology* **302,** 205-217 (2000).
41. Jordan, I.K., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research* **12,** 962-968 (2002).
42. Fang, G., Rocha, E. & Danchin, A. How Essential Are Nonessential Genes? *Molecular Biology and Evolution* **22,** 2147-2156 (2005).

43. Bruccoleri, R.E., Dougherty, T.J. & Davison, D.B. Concordance analysis of microbial genomes. *Nucleic Acids Research* **26**, 4482-4486 (1998).
44. Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., Han, S., Jeffery, L., Baek, S.-T., Lee, H., Shim, Y.S., Lee, M., Kim, L., Heo, K.-S., Noh, E.J., Lee, A.-R., Jang, Y.-J., Chung, K.-S., Choi, S.-J., Park, J.-Y., Park, Y., Kim, H.M., Park, S.-K., Park, H.-J., Kang, E.-J., Kim, H.B., Kang, H.-S., Park, H.-M., Kim, K., Song, K., Song, K.B., Nurse, P. & Hoe, K.-L. Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotech* **28**, 617-623 (2010).
45. Xu, P., Ge, X., Chen, L., Wang, X., Dou, Y., Xu, J.Z., Patel, J.R., Stone, V., Trinh, M., Evans, K., Kitten, T., Bonchev, D. & Buck, G.A. Genome-wide essential gene identification in *Streptococcus sanguinis*. *Sci. Rep.* **1**(2011).
46. Doyle, M., Gasser, R., Woodcroft, B., Hall, R. & Ralph, S. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11**, 222 (2010).
47. Zalacain, M., Biswas, S., Ingraham, K.A., Ambrad, J., Bryant, A., Chalker, A.F., Iordanescu, S., Fan, J., Fan, F., Lunsford, R.D., O'Dwyer, K., Palmer, L.M., So, C., Sylvester, D., Volker, C., Warren, P., McDevitt, D., Brown, J.R., Holmes, D.J. & Burnham, M.K.R. A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. *Journal of Molecular Microbiology and Biotechnology* **6**, 109-126 (2003).
48. Herce, H.D., Deng, W., Helma, J., Leonhardt, H. & Cardoso, M.C. Visualization and targeted disruption of protein interactions in living cells. *Nat Commun* **4**(2013).
49. Fields, S. & Song, O.-k. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245-246 (1989).
50. Cremazy, F.G.E., Manders, E.M.M., Bastiaens, P.I.H., Kramer, G., Hager, G.L., van Munster, E.B., Verschure, P.J., Gadella Jr, T.J. & van Driel, R. Imaging in situ protein–DNA interactions in the cell nucleus using FRET–FLIM. *Experimental Cell Research* **309**, 390-396 (2005).
51. Tsuganezawa, K., Nakagawa, Y., Kato, M., Taruya, S., Takahashi, F., Endoh, M., Utata, R., Mori, M., Ogawa, N., Honma, T., Yokoyama, S., Hashizume, Y., Aoki, M., Kasai, T., Kigawa, T., Kojima, H., Okabe, T., Nagano, T. & Tanaka, A. A Fluorescent-Based High-Throughput Screening Assay for Small Molecules That Inhibit the Interaction of MdmX with p53. *Journal of Biomolecular Screening* **18**, 191-198 (2013).
52. Acencio, M.L. & Lemke, N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* **10**, 290 (2009).
53. Hahn, M.W. & Kern, A.D. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Molecular Biology and Evolution* **22**, 803-806 (2005).
54. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genet* **2**, e88 (2006).
55. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).
56. Browne, F., Zheng, H., Wang, H. & Azuaje, F. From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions. *Advances in Artificial Intelligence* **2010**(2010).

57. Ebrahim, A., Lerman, J., Palsson, B. & Hyduke, D. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology* **7**, 74 (2013).
58. Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R. & Palsson, B.O. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protocols* **6**, 1290-1307 (2011).
59. Francke, C., Siezen, R.J. & Teusink, B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology* **13**, 550-558 (2005).
60. Lewis, N.E., Nagarajan, H. & Palsson, B.O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. *Nat Rev Micro* **10**, 291-305 (2012).
61. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. & Palsson, B.O. Reconstruction of biochemical networks in microorganisms. *Nat Rev Micro* **7**, 129-143 (2009).
62. Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B. & Stevens, R.L. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech* **28**, 977-982 (2010).
63. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
64. Schellenberger, J., Park, J., Conrad, T. & Palsson, B. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213 (2010).
65. Monk, J.M., Charusanti, P., Aziz, R.K., Lerman, J.A., Premyodhin, N., Orth, J.D., Feist, A.M. & Palsson, B.Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences* **110**, 20338-20343 (2013).
66. Thiele, I., Fleming, R.M.T., Bordbar, A., Schellenberger, J. & Palsson, B.Ø. Functional Characterization of Alternate Optimal Solutions of Escherichia coli's Transcriptional and Translational Machinery. *Biophysical Journal* **98**, 2072-2081 (2010).
67. Hyduke, D.R. & Palsson, B.Ø. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet* **11**, 297-307 (2010).
68. O'Brien, Edward J., Monk, Jonathan M. & Palsson, Bernhard O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161**, 971-987 (2015).
69. Orth, J.D., Thiele, I. & Palsson, B.O. What is flux balance analysis? *Nat Biotech* **28**, 245-248 (2010).
70. Basler, G. Computational Prediction of Essential Metabolic Genes Using Constraint-Based Approaches. in *Gene Essentiality*, Vol. 1279 (ed. Lu, L.J.) 183-204 (Springer New York, 2015).
71. Mahadevan, R., Edwards, J.S. & Doyle, F.J., 3rd. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys J* **83**, 1331-40 (2002).
72. Segrè, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* **99**, 15112-15117 (2002).

73. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 7695-7700 (2005).
74. Kim, P.-J., Lee, D.-Y., Kim, T.Y., Lee, K.H., Jeong, H., Lee, S.Y. & Park, S. Metabolite essentiality elucidates robustness of Escherichia coli metabolism. *Proceedings of the National Academy of Sciences* **104,** 13638-13642 (2007).
75. Zomorrodi, A.R., Suthers, P.F., Ranganathan, S. & Maranas, C.D. Mathematical optimization applications in metabolic networks. *Metabolic Engineering* **14,** 672-686 (2012).
76. Cheng, J., Xu, Z., Wu, W., Zhao, L., Li, X., Liu, Y. & Tao, S. Training Set Selection for the Prediction of Essential Genes. *PLoS ONE* **9,** e86805 (2014).
77. Domingos, P. A few useful things to know about machine learning. *Communications of the ACM* **55,** 78-87 (2012).
78. Lesko, L.J. & Woodcock, J. Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective. *Nat Rev Drug Discov* **3,** 763-769 (2004).
79. Chan, J.N.Y., Nislow, C. & Emili, A. Recent advances and method development for drug target identification. *Trends in Pharmacological Sciences* **31,** 82-88 (2010).
80. Jay, J. Fermented Foods and Related Products of Fermentation. in *Modern Food Microbiology* 371-409 (Springer Netherlands, 1992).
81. Breuer, U. Book Reviews: Microbial Production of Biopolymers and Polymer Precursors: Applications and Perspectives. Edited by Bernd H. A. Rehm. *CLEAN – Soil, Air, Water* **37,** 414-414 (2009).
82. Marshall, V.M. Probiotics and Prebiotics: Scientific Aspects (2005). *International Journal of Dairy Technology* **60,** 63-64 (2007).
83. Balter, S. Foodborne Pathogens: Microbiology and Molecular Biology. *Emerging Infectious Diseases* **12,** 2003-2003 (2006).
84. Hay, I.D., Rehman, Z.U., Moradali, M.F., Wang, Y. & Rehm, B.H.A. Microbial alginate production, modification and its applications. *Microbial Biotechnology* **6,** 637-650 (2013).
85. Lima, L.J.R., Kamphuis, H.J., Nout, M.J.R. & Zwietering, M.H. Microbiota of cocoa powder with particular reference to aerobic thermoresistant spore-formers. *Food Microbiology* **28,** 573-582 (2011).
86. Scheldeman, P., Herman, L., Foster, S. & Heyndrickx, M. Bacillus sporothermodurans and other highly heat-resistant spore formers in milk. *Journal of Applied Microbiology* **101,** 542-555 (2006).
87. Singh, S.N. & Tripathi, R.D. *Environmental bioremediation technologies*, xx, 518 p. (Springer, Berlin ; New York, 2007).
88. Pritchard, P.H., Mueller, J.G., Rogers, J.C., Kremer, F.V. & Glaser, J.A. Oil spill bioremediation: experiences, lessons and results from the Exxon Valdez oil spill in Alaska. *Biodegradation* **3,** 315-335 (1992).
89. Pieper, D.H. & Reineke, W. Engineering bacteria for bioremediation. *Current Opinion in Biotechnology* **11,** 262-270 (2000).
90. Dressaire, C., Gitton, C., Loubière, P., Monnet, V., Queinnec, I. & Cocaign-Bousquet, M. Transcriptome and Proteome Exploration to Model Translation Efficiency and Protein Stability in Lactococcus lactis. *PLoS Computational Biology* **5,** e1000606 (2009).
91. Alberch, P. From genes to phenotype: dynamical systems and evolvability. *Genetica* **84,** 5-11 (1991).

92. Tan-a-ram, P., Cardoso, T., Daveran-Mingot, M.-L., Kanchanatawee, S., Loubière, P., Girbal, L. & Cocaign-Bousquet, M. Assessment of the Diversity of Dairy Lactococcus lactis subsp. lactis Isolates by an Integrated Approach Combining Phenotypic, Genomic, and Transcriptomic Analyses. *Applied and Environmental Microbiology* **77**, 739-748 (2011).
93. Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, C.A., Smith, H.O. & Venter, J.C. Genome Transplantation in Bacteria: Changing One Species to Another. *Science* **317**, 632-638 (2007).
94. Kaur, B. & Chakraborty, D. Biotechnological and Molecular Approaches for Vanillin Production: a Review. *Applied Biochemistry and Biotechnology* **169**, 1353-1372 (2013).
95. Jansen, R., Greenbaum, D. & Gerstein, M. Relating Whole-Genome Expression Data with Protein-Protein Interactions. *Genome Research* **12**, 37-46 (2002).
96. Peng, C. & Gao, F. Protein localization analysis of essential genes in prokaryotes. *Sci Rep* **4**, 6001 (2014).
97. Goodacre, N.F., Gerloff, D.L. & Uetz, P. Protein Domains of Unknown Function Are Essential in Bacteria. *mBio* **5**(2014).
98. Lu, Y., Lu, Y., Deng, J., Lu, H. & Lu, L. Discovering Essential Domains in Essential Genes. in *Gene Essentiality*, Vol. 1279 (ed. Lu, L.J.) 235-245 (Springer New York, 2015).
99. Nelson, C., Hersh, B. & Carroll, S. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology* **5**, R25 (2004).
100. Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M.D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **34**, 564-574 (2006).
101. Rocha, E.P.C. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* **34**, 377-378 (2003).
102. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).

Chapter 3

From microbial gene essentiality to novel antimicrobial drug targets

Fredrick M. Mobegi
Sacha A. F. T. van Hijum
Peter Burghout
Hester J. Bootsma
Stefan P.W. de Vries
Christa E. Gaast-deJongh
Elles Simonetti
Jeroen Langereis
Peter W. M. Hermans
Marien I. de Jonge
Aldert Zomer

BMC Genomics 2014; 15(1): 958

Abstract

Background: Bacterial respiratory tract infections, mainly caused by *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Moraxella catarrhalis* are among the leading causes of global mortality and morbidity. Increased resistance of these pathogens to existing antibiotics necessitates the search for novel targets to develop potent antimicrobials.

Result: Here, we report a proof of concept study for the reliable identification of potential drug targets in these human respiratory pathogens by combining high-density transposon mutagenesis, high-throughput sequencing, and integrative genomics. Approximately 20% of all genes in these three species were essential for growth and viability, including 128 essential and conserved genes, part of 47 metabolic pathways. By comparing these essential genes to the human genome, and a database of genes from commensal human gut microbiota, we identified and excluded potential drug targets in respiratory tract pathogens that will have off-target effects in the host, or disrupt the natural host microbiota. We propose 249 potential drug targets, 67 of which are targets for 75 FDA-approved antimicrobials and 35 other researched small molecule inhibitors. Two out of four selected novel targets were experimentally validated, proofing the concept.

Conclusion: Here we have pioneered an attempt in systematically combining the power of high-density transposon mutagenesis, high-throughput sequencing, and integrative genomics to discover potential drug targets at genome-scale. By circumventing the time-consuming and expensive laboratory screens traditionally used to select potential drug targets, our approach provides an attractive alternative that could accelerate the much needed discovery of novel antimicrobials.

Background

The World Health Organization (WHO; www.who.int) ranks respiratory tract infections (RTI) among the ten leading causes of global mortality. RTI are associated with several bacterial species, of which *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis* are the most prevalent community-acquired respiratory bacterial pathogens [1]. In healthy individuals, these species colonize mucosal surfaces of the upper airways in a commensal state. Their relevance as pathogens arises when they infiltrate and colonize the otherwise sterile spaces in the middle ear, lung or bloodstream, progressing to disease [2]. With the mounting inexorable resistance of these pathogens against several commonly used antimicrobials [1], discovery of new protein targets against which new antibiotics could be developed will highly benefit global healthcare management of RTI.

Elucidation of genes essential for bacterial growth and viability is a prerequisite for identifying potential drug targets [3]. Essential genes are highly conserved and are thus considered as favourable drug targets for broad-spectrum inhibition [4]. On the other hand, some metabolic pathways constitute crucial transport and catalytic proteins which could also form attractive drug targets. Furthermore, most pathogens have drastically reduced their biosynthetic capabilities, and instead rely on their hosts to provide vital nutrients like amino acids, vitamins, and nucleobases [5]. Transport systems for these nutrients are generally conserved and indispensable for survival of the pathogen in its host [6], making them promising drug targets. In order to qualify as drug targets, microbial genes should meet several requirements. First, they should be nonhomologous to human genes to avoid drug cytotoxicity [3]. Additionally, targets should either be completely absent, or catalytically distinctive from genes found in host gut commensal microbiota, whose perturbation is likely to be detrimental to human nutrition, health, and physiology [7]. It has been shown that antibiotic killing of commensal microbiota facilitates proliferation, and often dominance, of antibiotic-resistant pathogens on mucosal surfaces [8]. Lastly, candidate drug targets must be accessible by inhibitors. Essential surface/membrane and secreted proteins are particularly promising, having been successfully targeted by protein drugs, and representing majority of all known drug targets [9, 10].

Previous microbial gene essentiality predictions employed techniques generally limited in specificity and/or throughput [11, 12]. These shortcomings are alleviated by high-throughput transposon insertion sequencing strategies, such as Tn-seq, TraDIS, INseq, or variants thereof, which have been applied in recent studies to comprehensively essay gene essentiality and genetic interactions in various bacteria [13, 14]. Here, we applied Tn-seq to reliably identify essential genes in *S. pneumoniae*, *H. influenzae* and *M. catarrhalis*. Products of these genes were compared against the human proteome, and the catalogue of genes from human gut commensal microbes, to identify and eliminate targets likely to have off-target effects in the host or on the host's gut microbiota. Two

out of four of the finally identified novel drug targets have been successfully validated using existing inhibitors. This study pioneers an integrative approach for rapid and cost-effective identification of novel drug targets. Our findings do not only improve the overall understanding of respiratory pathogens, but also serve as a proof of concept for the robust yet underexploited approaches, combining *in silico* and wet laboratory analyses in identifying antimicrobial drug targets, as recently reviewed [15]. This approach has allowed us to identify promising drug target leads, which after experimental validation could be potentially advanced to the discovery of novel antimicrobials for the treatment of RTI.

Methods

Bacterial genomes and gene reannotation

Whole genome sequences for *S. pneumoniae* TIGR4 uid57857, *S. pneumoniae* R6 uid57859, *H. influenzae* Rd KW20 uid57771, *H. influenzae* 86 028NP uid58093 and *M. catarrhalis* BBH18 uid48809 were obtained from the National Centre for Biotechnology Information (NCBI) Genbank File Transfer Protocol (FTP) website (ftp://ftp.ncbi.nih.gov/Genbank/). All open reading frame (ORF) annotations were updated using Rapid Annotation using Subsystem Technology (RAST) [16]. In this analysis, all locus coordinates in original Genbank genomes release were retained without adjustments for frame-shifts.

Orthology and gene essentiality predictions

We clustered the reannotated protein sequences into putative orthologous groups using the OrthoMCL standalone software Version 2.0.2 [17]. Most studies have consistently deciphered essential genes under ideal conditions, that is, in the richness of all necessary nutrients and without environmental stress. For the purpose of this study, we define the "essentiality" of a gene as its indispensability under rich media conditions, unless stated otherwise. The caveat with this approach is that essential genes required for metabolism within the host may be missed.

Transposon mutant libraries used were either created in-house for this study, or obtained from literature and reanalyzed. The *M. catarrhalis* BBH18 *marinerT7* transposon mutant libraries consisting of 28,000 and 7,000 independent transformants were previously described [18, 19], and the 12,500 transformants library was generated using the previously described protocol [18]. The 40,000 transformants *S. pneumoniae* R6 and the 11,000 transformants *H. influenzae* 86 028NP library were previously described [20, 21]. Libraries for the 15,000 transformants *S. pneumoniae* R6 and *H. influenzae* Rd KW20 were also respectively constructed as previously described [20, 21]. The Tn-seq technology was used to profile the relative abundance of each mutant in all libraries after growth as described previously [22], except for *S. pneumoniae* TIGR4. Tn-seq data for *S. pneumoniae*

TIGR4 were obtained from literature [23]. We then performed essentiality predictions for individual genes using the in-house developed web-tool, ESSENTIALS [24], which enabled us calculate a statistical essentiality metric for each ORF, and precisely delineate the optimal boundary between essential and nonessential ORFs in each of the 5 strains. Analysis data can be found at http://bamics2.cmbi.ru.nl/websoftware/essentials/links.html

Overrepresented metabolic pathways and subsystems

Pathways and subsystems for the strains under study were obtained from the Kyoto Encyclopedia of Genes and Genomes orthology, and the SEED databases respectively [25, 26]. Using a Fisher's exact test, we performed functional categories enrichment for the pathways and subsystems, while incorporating the statistical essentiality value (the fold-change value predicted by ESSENTIALS) for each ORF. We corrected for multiple testing using Bonferroni correction and obtained *q*-values for corresponding *p*-values [27].

Proteins subcellular localization (SCL)

The subcellular localizations (SCL) of all proteins in this study were determined using publicly available SCL prediction tools. First, we analyzed all Gram-positive and Gram-negative strains using pSORTdb version 2.0 [10] and CELLO version 2.5 [28]. Further complementation SCL predictions were performed using LocateP and GnegPloc for Gram-positive and Gram-negative strains respectively [29, 30]. Additionally, the presence of integral Gram-negative outer membrane proteins (OMP) was determined using β-barrel outer membrane protein predictor (BOMP) [31]. Proteins that showed different SCL predictions in the different predictors used were denoted "Unknown", together with those predicted to be of unknown SCL by majority of the predictors used.

Selecting potential drug targets

To identify and eliminate essential genes with close undesirable orthologs, we performed separate unidirectional blast (BLASTP) searches, using an *E*-value cut-off of 1e-10, and minimum 70% sequence identity over 75% sequence coverage; against the human genome, and the metagenomics catalogue of non-redundant human gut microbiome genes by Qin *et al* [7].

Determination of antimicrobial activity

Selection of potential drug targets for *in vivo* validation was mainly based on their novelty, that is, they have not been described as targets to existing antimicrobials. Commercial availability of inhibitory compounds without resorting to customized chemical synthesis was also key; all inhibitory compounds used were supplied by Sigma Aldrich. 1-Methyluric acid, 5, 5'-Dithio-bis-(2-nitrobenzoic Acid), and 5'-deoxyadenosine were dissolved in water at 5mg/ml. When necessary, the pH was neutralized (to pH7) using 10M NaOH solution or 1M HCl. Antimicrobial activity of the compounds was tested

by Kirby-Bauer/disk diffusion assay [32], by applying 10 µg of the inhibitory compounds to 6mm filter paper discs at concentration ranging from 10000 to 0.05 µg/ml in 10-fold stepwise dilutions. As for (R)-6-fluoromevalonate diphosphate 2 µl of (R)-6-fluoromevalonate diphosphate was diluted in 1ml of Milli-Q (MQ). 10 µl and 100 µl of the dilution were used in separate disk diffusion assays. Columbia III agar with 5% sheep blood medium was used for *S. pneumoniae*. Brain heart infusion (BHI) agar medium and a combination medium of BHI, hemin, and NAD were used for *M. catarrhalis* and *H. influenzae* respectively. MIC calculations were performed as described by Wiegand and colleagues [33]. Experiments were performed in quadruplicate, and outliers were removed using the Grubbs test [34].

Toxicity assays on epithelial cell lines

Cellular toxicity of (R)-6-fluoromevalonate diphosphate was tested using the CellTox Green Cytotoxicity Assay (Promega, WI) on Detroit 562 (ATCC CCL-138) and A549 (ATCC CCL-185) cell lines according to the manufacturer's instructions. The two cell lines were exposed to (R)-6-fluoromevalonate diphosphate at its effective MIC concentration of 26.6 µg/ml for 24 hours at 37°C with 5 % CO_2. Fluorescence was measured on a Perkin Elmer 1420 Victor 3V multi-label plate reader.

Results and discussion

Genome reannotation and gene clustering

We sought to determine potential drug targets in *S. pneumoniae*, *H. influenzae*, and *M. catarrhalis* following the selection criteria outlined in Figure 1. For these species, five strains with the required Tn-seq data were available; *S. pneumoniae* strains R6 and TIGR4, *H. influenzae* strains Rd KW20 and 86 028NP, and *M. catarrhalis* strain BBH18. Altogether, genomes of these strains in their initial annotations constituted of 10,072 open reading frames (ORFs). These annotations were updated using RAST to ensure consistency and comparability among strains in subsequent analyses. This analysis resulted in putative annotations for about 50% of all ORFs originally annotated with a hypothetical function (Table 1; Additional file 1). Next, we clustered the updated protein sequences using OrthoMCL, producing 1,798 orthologous groups/clusters (OGs) with, and 2,729 without singletons respectively (Additional file 1). This clustering of orthologous proteins allowed for the determination of species and/or strain specific proteins, as well as determining the metabolic potential of the strains. For example, the "Gram-negative specific" periplasmic chaperones (SurA) were clustered in OG_756 (cluster 756), while the "Streptococci-specific" transcriptional regulators (LytR) were clustered in OG_2554. On the other hand, 300 OGs, including OG_184, OG_186, OG_216, and OG_224, among others, contained genes conserved in all the five strains. All protein in individual OGs constituted of similar

or identical functional annotations. This consistency in grouping and annotation was observed across all OGs, suggesting a reliable clustering. Confirmatory clusters and respective annotations derived from the clusters of orthologous genes (COGs) database were consistent with our OrthoMCL clusters. Additionally, using the OG's, we were able to curate annotations for the HI1586 locus in *Haemophilus influenzae* Rd KW20, which was possibly misannotated in the initial release, as an isoleucyl-tRNA synthetase instead of a Na^+/H^+ antiporter.

Essential and conserved protein-coding genes

Loss of mutant readouts from a transposon library after *in vitro* transposition and genetic transformation of the wild-type isolate is a strong indicator of gene essentiality [35]. Although some essential genes tolerate disruptive insertions in the 3' regions, generally, insertions in essential genes lead to lethal phenotypes [36]. For our analysis, mutant libraries and/or Tn-seq data were constructed in in-house experiments or obtained from literature (Table 1). We separately analyzed the Tn-seq datasets using ESSENTIALS [24].

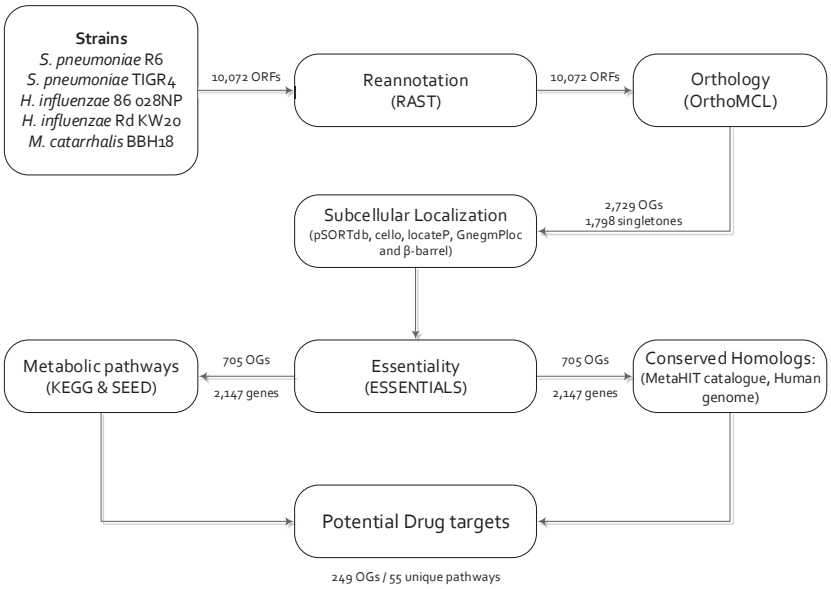


Figure 1. Schematic overview of the drug target selection criteria**.** Genome annotations information for *S. pneumoniae* R6, *S. pneumoniae* TIGR4, *H. influenzae* 86 028NP, *H. influenzae* Rd KW20, and *M. catarrhalis* BBH18 were updated using RAST. The proteins with updated annotations were then clustered into putative orthologous groups using OrthoMCL, and their subcellular localizations predicted in various publicly available tools. ESSENTIALS was used to analyze various transposon mutant libraries and predict the essentiality metric for each ORF. Comparing the ensuing essential genes with the catalogue of human gut microbial genes, as well as with the human genome helped to eliminate genes with conserved orthologs, and subsequently prioritize potential drug targets.

This analysis resulted in the identification of 532 essential genes in *H. influenzae* 86-028NP, representing 28% of the genome, a higher number as compared to the other Gram-negative strains; *H. influenzae* Rd KW20 and *M. catarrhalis* BBH18, in which we identified 431 and 445 essential genes respectively. In *S. pneumonia*, we identified 325 and 414 essential genes for the R6 and TIGR4 strains respectively (Table 1; Additional file 1). These values showed that on average, about 20% of all genes in the five strains are essential. This is consistent with earlier studies which have reported 15-25% of all genes in a genome being essential [23, 36, 37].

Differences in the number of essential genes could be explained by various factors that hamper precision in transposon mutagenesis experiments, including short gene lengths and unsaturated transposon libraries; "saturation" being the presence of at least one insertion in every gene. In practice, short genes are less susceptible to disruptive transposon insertions, hence, more likely to be misclassified as essential. In unsaturated transposon mutant libraries, dispensable genes are also more likely to be devoid of transposon insertions, and therefore misclassified as being essential genes. The low-density transposon mutant library (approximately 11,000 colony forming units; CFU) used for *H. influenzae* 86-028NP, and a substantial number of short genes in its genome could, therefore, explain the apparently overestimated (532) essential genes. Relatively saturated libraries of approximately 20,000 CFU and 40,000 CFU were used for *H. influenzae* Rd KW20 and *M. catarrhalis* BBH18 respectively (Table 1). A rarefaction analysis on our data confirmed that the *S. pneumoniae*, *M. catarrhalis*, and *H. influenzae* Rd KW20 transposon libraries approached saturation (Additional file 4). Additionally, based on derivations of Poisson's law, there is a 99.6% probability that genes with a size of 1kb are hit in the 1.9Mb *H. influenzae* 86-028NP genome and an 11,000 CFU mutant library. Similar statistics on the 1.79Mb *H. influenzae* Rd KW20 genome with a 20,000 CFU mutant library shows a 99.99% probability. Therefore, *H. influenzae* 86-028NP could have suffered slightly more false positive predictions due to its less saturated mutant libraries.

We selected 705 OGs containing at least one essential gene from any of the five strains for further analysis. These essential OGs mainly consist of proteins with annotated functions, participating in diverse core cellular processes, such as DNA replication, DNA transcription, protein translation, cell wall biosynthesis, signal transduction, and metabolism. Eighteen OGs, however, contained conserved proteins of uncharacterized function (Additional file 1). Functional characterization of these genes will aid in achieving the optimal set of targets that can be used to develop antimicrobials against RTI causing bacteria. The distribution and overlap of the essential genes within the three species is outlined in Figure 2. From the 705 OGs, we identified 128 OGs that constituted of genes conserved and essential in all five trains, representing targets particularly attractive for developing broad-spectrum antimicrobials to treat RTI, since they encode components of basal cellular functions in respiratory pathogens. Importantly, collective analysis of the five strains revealed species-specific and/or "Gram-category" specific essential genes, best suited for narrow-spectrum or specialized inhibition.

Table 1: Strain genome annotation updates and essentiality predictions. Transposon mutant libraries and Tn-seq data prepared for this study (*), or Tn-seq data sequenced in this study from mutant libraries obtained from literature (**); otherwise, all data was obtained from literature and reanalyzed in this study.

			Annotations update		Essentiality predictions				
Strain	Genbank accession	Total number of ORFs	ORFs with hypothetical function in genome	ORFs with hypothetical function after RAST	Number of insertion sites [a]	Log_2 fold change cutoff [b]	Mutant library size (CFU)	Number of sequenced reads [c]	Total essential genes
S. pneumoniae R6	NC003098	2,116	735	362	133,135	-6.45	40,000 15,000*	8,906,301 4,400,836** 5,641,892* 6,335,218*	325
S. pneumoniae TIGR4	NC003028	2,302	738	458	141,459	-4.43	6 x 20,000	876,181 855,535 825,675 1,294,187 1,241,843 1,291,425	414
H. Influenzae 86 028NP	NC007146	1,900	456	233	138,229	-4.64	11,000	5,751,765 4,880,492 9,925,569 9,517,400	532
H. influenzae Rd KW20	NC000907	1,790	429	118	131,955	-4.59	20,000*	3,857,040* 3,229,286* 8,152,867* 7,724,536*	431
M. catarrhalis BBH18	NC014147	1,964	586	573	116,242	-4.70	28,000 12,500* 7,000	3,522,998** 4,618,913* 4,697,209	445

[a] Total number of possible unique transposon insertion sites in the genome; [b] the computed fold change cut-off that separates essential and nonessential genes in each strain; [c] number of sequence reads generated by the Illumina HiSeq sequencer.

Essential metabolic pathways and subsystems

Functional category enrichment analyses were performed for all KEGG metabolic pathways and the SEED subsystems [25, 26]. As of August 22, 2013, the KEGG database describes 448 fully characterized pathways, which are further subcategorized into 262,304 reference maps for various organisms. All KEGG characterized proteins in the 705 essential OGs could be assigned to 84 unique pathways. Among these, characterized proteins contained in the 128 OGs that are conserved and essential in all five strains could be assigned to 47 metabolic pathways (Additional files 1 and 2). As was the case for essential genes, the identified essential pathways specify among other functions, core bacterial bioprocesses like membrane transport, DNA replication and repair, signal transduction, metabolism, transcription and translation, ribosomal functions, and cellular processes including cell motility. The SEED, an alternative to KEGG, comprehensively groups genes at the level of a biological system and its subsystems. Currently, there are approximately 1,009 characterized SEED subsystems. Use of SEED subsystems on the essential OGs also revealed overrepresentation of critical system processes, including those involved in protein biosynthesis, virulence, disease and defense, as well as metabolism of cofactors, vitamins, prosthetic groups, pigments, fatty acids, lipids, and isoprenoids (Table 2; Additional file 3).

Protein subcellular localization

Out of the 705 OGs selected, the majority (526) consists of cytoplasmic proteins. Cellular localization of the other OGs were predicted to be: 96 in the inner membrane, 11 in the outer membrane, 12 in the periplasm, and 4 in the extracellular space. In addition, 21 OGs are non-categorically predicted to contain membrane proteins, whereas 35 are of unknown localizations. Of the 11 outer membrane OGs, 7 contained β-barrels (Additional file 1).

Orthologs in human and human gut microflora

The human gut is home to microbiota whose proper composition and functioning collectively influence human nutrition, protection against pathogens and development of disease [7]. Perturbing this microbiota with antibiotics could cause adverse side effects. Furthermore, interference with human cell physiology by antibiotics as a consequence of non-specific targeting can cause severe cellular cytotoxicity [3], which may result in organ failure or even death. We used BLASTP analyses against the human genome (Genome Reference Consortium) and the human gut microbial gene catalogue [7], to identify targets that would likely have off-target effects. It is noteworthy that targets with as few as 10 matches in the non-redundant gut microbial gene catalogue were allowed in the final selection, as we hypothesized that these would have no effects on the gut microbiome preventing disruption of gut health. This decision was motivated by the

observation from our analysis that well known targets for both clinically approved antimicrobials and experimental small molecule inhibitors collated in DrugBank (Additional file 1; column 9) maintained on average fewer than 10 blast hits against the human gut microbial gene catalogue (Additional file 1; column 20). On the other hand, the majority of the targets with numerous blast hits were aminoacyl-tRNA synthetases (aaRSs) and ribosomal protein, including rpsL, a well-known target that had 249 hits for pneumococci, 156 for *H. influenzae*, and 151 for *M. catarrhalis*. One shortcoming of using such filtering criteria is that novel targets that have more than 10 blast hits are not effectively retained in the final selection. Nevertheless, we identified 96 OGs with orthologs in human, and 127 OGs with orthologs in human gut microflora, that is, with >10 blast hits (Additional file 1). All 20 aminoacyl-tRNA synthetases (aaRSs), essential for protein synthesis, were particularly conserved in both human and human gut microflora. Studies have shown that aaRSs can be selectively targeted as most bacterial aaRSs recognize and aminoacylate only cognate tRNA [38]. However, possible side effects are expected from drugs targeting aaRSs. RNA molecules and ribosomal proteins were also highly conserved in gut microbiota and humans. Additionally, the relatively short lengths and the presence of highly repetitive DNA in RNA sequences also rendered their essentiality predictions unreliable. All these molecules were therefore not included in the final selection of drug targets. Moreover, blast comparison between finally selected targets and their human orthologs showed minimal sequence identities (<35%) over short sequence coverage.

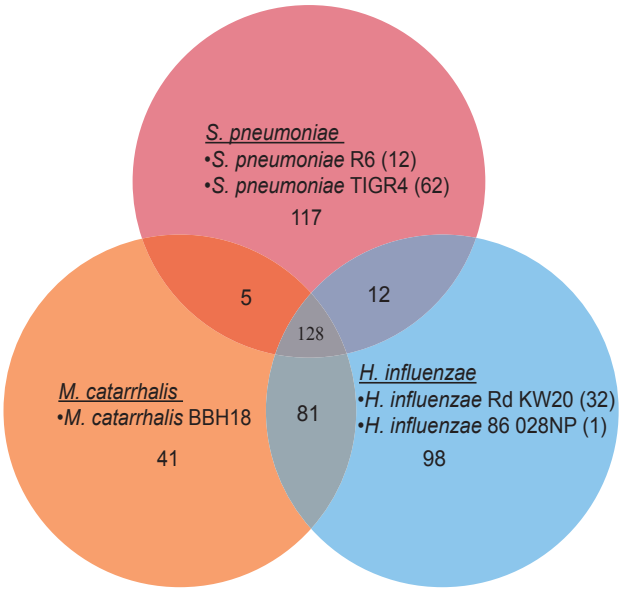


Figure 2. A Venn diagram showing the overlap of essential orthologous groups among the respiratory pathogens. Singletons are shown in brackets.

Table 2: Distribution of essential features among respiratory pathogens.

	Quantity of features in the strains				
	mct	hin	hit	spn	spr
Essential structural and non-coding RNAs	**5**	**49**	**41**	**136**	**47**
tRNA	4	18	0	12	8
rRNA	1	31	41	44	30
sRNA	n/a	n/a	n/a	80	9
Essential Protein-coding ORFs	**445**	**431**	**532**	**414**	**325**
Protein of unknown functions	159	172	225	186	127
Metabolism	173	142	182	124	100
Genetic Information Processing	93	95	101	95	93
Environmental Information Processing	20	24	24	9	5
Overrepresented/essential KEGG pathways	**236**	**437**	**196**	**307**	**356**
Metabolism	136	221	95	171	213
Genetic Information Processing	74	177	74	128	129
Environmental Information Processing	26	38	26	8	14
Cellular Processes	0	1	1	0	0
Overrepresented/essential SEED subsystems	**449**	**513**	**602**	**450**	**355**
Protein metabolism	84	85	99	100	93
Cofactors, Vitamins, Prosthetic Groups, Pigments	75	61	80	29	25
Cell Wall and Capsule	47	60	78	47	30
Amino Acids and Derivatives	41	59	58	14	11
Respiration	41	16	34	8	7
Fatty Acids, Lipids, and Isoprenoids	29	36	40	26	21
RNA Metabolism	25	59	71	60	39
Carbohydrates	24	30	46	47	35
DNA Metabolism	19	37	35	45	41
Stress Response	18	17	9	10	8
Nucleosides and Nucleotides biosynthesis	17	13	11	25	9
Virulence, Disease and Defense	16	18	18	16	15
Regulation and Cell signaling	8	4	8	6	5
Cell Division and Cell Cycle	5	18	15	17	16

The strains under study are abbreviated: mct; *Moraxella catarrhalis* BBH18, hin; *Haemophilus influenzae* Rd KW20, hit; *H. influenzae* 86 028NP, spn; *Streptococcus pneumoniae* TIGR4, and spr; *S. pneumoniae* R6. Untested categories are denoted by "n/a".

Drug targets selection and validation

We identified 249 potential drug targets in the five strains (Additional file 5), including key enzymes in pathways such as fatty acid biosynthesis [39-41], vitamin biosynthesis [42-45], and isoprenoid biosynthesis pathways [46-48], which have gained interest in drug discovery research, as well as 67 known targets inhibited by 75 FDA-approved antimicrobial drugs and 35 other researched small molecule inhibitors collated in the DrugBank database [49]. To validate our target prediction, we selected four novel targets with commercially available novel inhibitors of their predicted essential functions, that is, inhibitors not yet approved as clinical drugs and don't require to be custom synthesized: We tested whether exposure to these compounds inhibited growth of the target organisms.

Vitamin biosynthetic pathways constitute an attractive and largely untapped source of potential drug targets [42, 45]. For instance, thiamine (vitamin B1) in its active form

thiamine diphosphate is indispensable for the activity of the carbohydrate and branched-chain amino acid metabolic enzymes [42]. Most bacteria synthesize thiamine *de novo,* whereas humans depend on dietary uptake, making the thiamine biosynthetic pathway an attractive selective drug target. Folic acid (vitamin B9) is another indispensable cofactor, whose biosynthetic pathway was a target for sulfamidochrysoidine (prontosil); later replaced by an improved sulphonamide drug sulfanilamide, the first ever antibiotic used in humans [50]. The pathway is also targeted by trimethoprim [45], another clinically acknowledged chemotherapeutic agent that acts on dihydrofolate reductase. Niacin (vitamin B3; alternatively known as nicotinamide or nicotinic acid) is also essential to all living cells and is biosynthetically converted to nicotinamide adenine dinucleotide (NAD^+), a coenzyme involved in electron transport reactions in cell metabolism processes [51]. After it was described that niacin has therapeutic effects and it modulates various biological effects as well as NAD^+ metabolism, there has been an increased interest in the role of NAD^+ biosynthetic pathway in health and disease [52]. Prospects of targeting the pathway are also being explored. We used 5, 5'-Dithiobis, 2-nitrobenzoic acid (CAS: 69-78-3) to inhibit NAD^+ kinase (EC: 2.7.1.23) - a key enzyme in the NADP biosynthesis which catalyzes the phosphorylation of NAD^+ into $NADP^+$. Inhibition of growth was expected in both Gram-positive and Gram-negative strains. However, inhibition of growth was only observed in the disk diffusion assays for *S. pneumoniae*. MIC calculations were inconclusive as they ranged from 319 to 2500 µg/ml with a large variability between assays (Table 3).

As an essential amino acid, methionine is not synthesized *de novo* in humans, who must rely on dietary intake. Enzymes involved in microbial methionine biosynthesis therefore offer highly specific and selective drug targets. We used 1-methyluric acid (CAS: 708-79-2) to target S-adenosylmethionine synthetase (EC: 2.5.1.6); a key enzyme in methionine biosynthesis, whose drug target potential has been explored in various pathogens [53, 54]. Contrary to expectations, no growth inhibition was observed in Gram-negative strains (Table 3; Figure 3): growth inhibition was only observed in *S. pneumoniae*. Since 1-methyluric acid formed a precipitate in concentrations above 312 µg /ml, no MIC values could be calculated. This lack of growth inhibition in Gram-negative strains may possibly be due to their double layered cell walls which are less penetrable [55], or the bacteria have expanded their resistance mechanisms to evade killing by antimicrobials [55, 56]. It is also possible that the two Gram-negative species have alternative mechanisms for methionine biosynthesis, further complicating screening for effective drugs.

The microbial fatty acid synthesis (FAS) pathway is an attractive target for drug discovery [41, 57]. This pathway is subdivided into type I and II, whereby human FAS proteins predominantly belong to type I FAS, and the bacterial ones are predominantly type II FAS. Proteins from the two FAS types generally possess distinctive molecular organization of the active site allowing for selective targeting [39, 40]. Although Gram-positive pathogens could compensate FASII inhibition by assimilating environmental fatty acids; particularly unsaturated fatty acids [58, 59], several clinical and household

antimicrobials targeting key FAS enzymes, e.g. Platensimycin and Platencin have been successfully developed [41, 60].

In our analysis, we identified various genes conserved in all five strains, for example genes in OGs 085, 143, and 653, whose products play key roles in the FAS pathway. With 5'-Deoxyadenosine (CAS: 4754-39-6), we targeted lipoate synthase (LipA; EC: 2.8.1.8), a key enzyme in the lipoic acid metabolism [61], using product-level inhibition. Surprisingly, we observed growth inhibition in all three species (Figure 3; Table 3), despite the target cluster (OG_653) comprising of orthologs from only Gram-negative strains (Additional file 1). This observations are also reflected in the MIC, which ranged from 29.3 to 205.1 µg/ml (Table 3). A blastP comparison showed that the closest ortholog of the Gram-negative LipA in *S. pneumoniae* is the non-lipoic pathway enzyme fructose-6-phosphate aldolase I, sharing about 32% sequence identity. Moreover, a comparison between LipA and lipoate-protein ligase (LplA), the key lipoylation enzyme in *S. pneumoniae* [61], revealed that the two proteins are non-orthologous, as they share very low sequence identity (<25%). They however have conserved domain which may explain the observed growth inhibition.

Table 3. Drug target *in vivo* validation summary. Diameter of the clearance zone after normal incubation represents the inhibition area on disk. Concentrations showing delayed growth are denoted by an asterisk (*).

Compound	Amount on disc (µg)	MIC µg/ml; Std. Dev. [Inhibition area on disk diffusion assay]		
		S. pneumoniae	*H. influenzae*	*M. catarrhalis*
5,5'-dithiobis(2-nitrobenzoate) (CAS 69-78-3)	1,000	2,500; 0 [4 mm*]	781; 313 [none]	319; 303 [none]
1-methyluric acid (CAS 708-79-2)	1,000	>312.5 [6 mm]	>312.5 [none]	>312.5 [none]
5'deoxyadenosine (CAS 4754-39-6)	1,000	78.1; 0 [6 mm]	205; 132 [5 mm*]	29.3; 11 [12 mm]
(R)-6-fluoromevalonate diphosphate (CAS 2822-77-7)	1,000	26.6; 11.5 [12 mm]	4,167; 1443 [none]	>5,000; 0 [none]
(R)-6-fluoromevalonate diphosphate (CAS 2822-77-7)	100	26.6; 11.5 [4 mm*]	4,167; 1443 [none]	>5,000; 0 [none]

Std. Dev. = Standard deviation

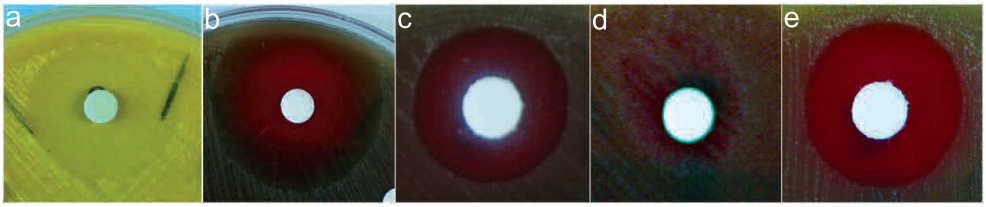


Figure 3. Validation of growth inhibition using disk diffusion essays. Cell culture plate cross-sectional images showing the area of growth inhibition for: a. *M. catarrhalis* in 5'deoxyadenosine, and *S. pneumoniae* in; b. (R)-6-fluoromevalonate diphosphate, c. I methyl, d. 5, 5'-dithiobis (2-nitrobenzoate), and e. 5'deoxyadenosine respectively.

Isoprenoids are natural products involved in many biochemical functions, such as supplying quinones for the electron transport chains, components of membranes, and subcellular targeting and regulation [47]. Humans employ the mevalonate pathway, whereas most microbes follow a non-mevalonate (1-deoxy-d-xylulose 5-phosphate/2-C-methyl-d-erythritol 4-phosphate) pathway. Functional roles of key enzymes in the isoprenoid biosynthesis pathway are well characterized, opening prospects for the discovery of novel drug targets [46, 48]. Fosmidomycin is a promising isoprenoid-based anti-malarial drug which is currently in clinical trials [48]. Using 6-fluoromevalonate (CAS: 2822-77-7) to target diphosphomevalonate decarboxylase (EC: 4.1.1.33), we observed selective growth inhibition only in *S. pneumoniae* as expected (Additional file 1; Figure 3; Table 3). Additionally, no effects on growth were observed in the Gram-negative strains, which was also as expected. We determined an average MIC of value 26.6 µg/ml for the *S. pneumoniae* growth inhibition (Table 3). At 26.6 µg/ml, no toxicity was observed in cell toxicity assays on epithelial cell lines (data not shown). Moreover, in patent WO 1995013058 A1, no cytotoxic effects of 6-fluoromevalonate were observed on T-lymphocytes. Previous literature also shown that 6-fluoromevalonate could potentially function the same as statins, as they inhibit the same pathway [62]. Diphosphomevalonate decarboxylase could therefore be a promising target for developing novel antibiotics against *S. pneumoniae* [63].

Conclusion

We have combined Tn-seq with *in silico* approaches to obtain an insight into many essential and conserved molecular functions, which we predicted to be unique among respiratory pathogens. With this combinatorial approach, we have reliably identified 249 potential drug targets, 67 of which are acknowledged targets for 75 FDA-approved antimicrobial drugs and 35 other researched small molecule inhibitors [49]; we successfully validated two of the four tested targets. Here, we propose a number of novel potential drug targets that are a concrete lead for experimental validation. We anticipate that future research based on this study will eventually provide interesting targets that can be successfully moved to drug development. In conclusion, we have pioneered a powerful approach, which combines gene essentiality data with robust computational techniques, to comprehensively screen for antimicrobial drug targets at genome-scale. This approach circumvents the complex and costly laboratory screens, thus, facilitating directed drugs discovery.

Supporting data

The datasets supporting the results of this article are included within the article and its additional files online at http://www.biomedcentral.com/1471-2164/15/958/additional. Tn-Seq datasets are available in the European Nucleotide Archive repository at http://www.ebi.ac.uk/ena/data/view/PRJEB7553.

Acknowledgements

This work was supported by funding from the European Commission FP7 Marie Curie IEF Action [274586 to AZ] and the Netherlands Genomics Initiative Horizon Breakthrough [93518023 to PB].

References

1. Hoban, D.J., Doern, G.V., Fluit, A.C., Roussel-Delvallez, M. & Jones, R.N. Worldwide prevalence of antimicrobial resistance in *Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis* in the SENTRY Antimicrobial Surveillance Program, 1997-1999. *Clin Infect Dis* **32 Suppl 2**, S81-93 (2001).
2. Lijek, R.S. & Weiser, J.N. Co-infection subverts mucosal immunity in the upper respiratory tract. *Curr Opin Immunol* **24**, 417-23 (2012).
3. Duffield, M., Cooper, I., McAlister, E., Bayliss, M., Ford, D. & Oyston, P. Predicting conserved essential genes in bacteria: *in silico* identification of putative drug targets. *Mol Biosyst* **6**, 2482-9 (2010).
4. Sakharkar, K.R., Sakharkar, M.K. & Chow, V.T. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico Biol* **4**, 355-60 (2004).
5. Lewis, K. Multidrug resistance: Versatile drug sensors of bacterial cells. *Curr Biol* **9**, R403-7 (1999).
6. Clayton, R.A., White, O., Ketchum, K.A. & Venter, J.C. The first genome from the third domain of life. *Nature* **387**, 459-62 (1997).
7. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Meta, H.I.T.C., Bork, P., Ehrlich, S.D. & Wang, J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
8. Buffie, C.G. & Pamer, E.G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* **13**, 790-801 (2013).
9. Ahram, M. & Springer, D.L. Large-scale proteomic analysis of membrane proteins. *Expert Rev Proteomics* **1**, 293-302 (2004).
10. Yu, N.Y., Laird, M.R., Spencer, C. & Brinkman, F.S. PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res* **39**, D241-4 (2011).
11. Molzen, T.E., Burghout, P., Bootsma, H.J., Brandt, C.T., van der Gaast-de Jongh, C.E., Eleveld, M.J., Verbeek, M.M., Frimodt-Moller, N., Ostergaard, C. & Hermans, P.W. Genome-wide identification of *Streptococcus pneumoniae* genes essential for bacterial replication during experimental meningitis. *Infect Immun* **79**, 288-97 (2011).
12. Sassetti, C.M., Boyd, D.H. & Rubin, E.J. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci U S A* **98**, 12712-7 (2001).
13. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* **11**, 435-42 (2013).
14. Barquist, L., Boinett, C.J. & Cain, A.K. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biology* **10**, 1161-1169 (2013).
15. Chung, B.K.-S., Dick, T. & Lee, D.-Y. In silico analyses for the discovery of tuberculosis drug targets. *Journal of Antimicrobial Chemotherapy* **68**, 2701-2709 (2013).
16. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L.,

Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).
17. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13,** 2178-89 (2003).
18. de Vries, S.P.W., Burghout, P., Langereis, J.D., Zomer, A., Hermans, P.W.M. & Bootsma, H.J. Genetic requirements for *Moraxella catarrhalis* growth under iron-limiting conditions. *Molecular Microbiology* **87,** 14-29 (2013).
19. de Vries, S.P., Eleveld, M.J., Hermans, P.W. & Bootsma, H.J. Characterization of the molecular interplay between *Moraxella catarrhalis* and human respiratory tract epithelial cells. *PLoS One* **8,** e72193 (2013).
20. Burghout, P., Cron, L.E., Gradstedt, H., Quintero, B., Simonetti, E., Bijlsma, J.J.E., Bootsma, H.J. & Hermans, P.W.M. Carbonic Anhydrase Is Essential for *Streptococcus pneumoniae* Growth in Environmental Ambient Air. *Journal of Bacteriology* **192,** 4054-4062 (2010).
21. Langereis, J.D., Zomer, A., Stunnenberg, H.G., Burghout, P. & Hermans, P.W.M. Nontypeable *Haemophilus influenzae* Carbonic Anhydrase Is Important for Environmental and Intracellular Survival. *Journal of Bacteriology* **195,** 2737-2746 (2013).
22. Burghout, P., Zomer, A., van der Gaast-de Jongh, C.E., Janssen-Megens, E.M., Françoijs, K.-J., Stunnenberg, H.G. & Hermans, P.W.M. *Streptococcus pneumoniae* Folate Biosynthesis Responds to Environmental CO2 Levels. *Journal of Bacteriology* **195,** 1573-1582 (2013).
23. van Opijnen, T., Bodi, K.L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6,** 767-72 (2009).
24. Zomer, A., Burghout, P., Bootsma, H.J., Hermans, P.W. & van Hijum, S.A. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One* **7,** e43012 (2012).
25. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40,** D109-14 (2012).
26. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33,** 5691-702 (2005).
27. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100,** 9440-5 (2003).
28. Yu, C.S., Chen, Y.C., Lu, C.H. & Hwang, J.K. Prediction of protein subcellular localization. *Proteins* **64,** 643-51 (2006).
29. Zhou, M., Boekhorst, J., Francke, C. & Siezen, R.J. LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics* **9,** 173 (2008).
30. Shen, H.B. & Chou, K.C. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J Theor Biol* **264,** 326-33 (2010).

31. Berven, F.S., Flikka, K., Jensen, H.B. & Eidhammer, I. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res* **32,** W394-9 (2004).
32. Bauer, A.W., Perry, D.M. & Kirby, W.M. Single-disk antibiotic-sensitivity testing of staphylococci: An analysis of technique and results. *A.M.A. Archives of Internal Medicine* **104,** 208-216 (1959).
33. Wiegand, I., Hilpert, K. & Hancock, R.E. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc* **3,** 163-75 (2008).
34. Grubbs, F.E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11,** 1-21 (1969).
35. Gawronski, J.D., Wong, S.M., Giannoukos, G., Ward, D.V. & Akerley, B.J. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A* **106,** 16422-7 (2009).
36. Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H. & Shapiro, L. The essential genome of a bacterium. *Mol Syst Biol* **7,** 528 (2011).
37. Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N. & Mekalanos, J.J. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci U S A* **99,** 966-71 (2002).
38. Ochsner, U.A., Sun, X., Jarvis, T., Critchley, I. & Janjic, N. Aminoacyl-tRNA synthetases: essential and still promising targets for new anti-infective agents. *Expert Opin Investig Drugs* **16,** 573-93 (2007).
39. Campbell, J.W. & Cronan, J.E., Jr. Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery. *Annu Rev Microbiol* **55,** 305-32 (2001).
40. Payne, D.J., Warren, P.V., Holmes, D.J., Ji, Y. & Lonsdale, J.T. Bacterial fatty-acid biosynthesis: a genomics-driven target for antibacterial drug discovery. *Drug Discov Today* **6,** 537-544 (2001).
41. Manallack, D.T., Crosby, I.T., Khakham, Y. & Capuano, B. Platensimycin: a promising antimicrobial targeting fatty acid synthesis. *Curr Med Chem* **15,** 705-10 (2008).
42. Du, Q., Wang, H. & Xie, J. Thiamin (vitamin B1) biosynthesis and regulation: a rich source of antimicrobial drug targets? *Int J Biol Sci* **7,** 41-52 (2011).
43. Debnath, J., Siricilla, S., Wan, B., Crick, D.C., Lenaerts, A.J., Franzblau, S.G. & Kurosu, M. Discovery of selective menaquinone biosynthesis inhibitors against *Mycobacterium tuberculosis*. *J Med Chem* **55,** 3739-55 (2012).
44. Kronenberger, T., Schettert, I. & Wrenger, C. Targeting the vitamin biosynthesis pathways for the treatment of malaria. *Future Med Chem* **5,** 769-79 (2013).
45. Bermingham, A. & Derrick, J.P. The folic acid biosynthesis pathway in bacteria: evaluation of potential for antibacterial drug discovery. *Bioessays* **24,** 637-48 (2002).
46. Dhar, M.K., Koul, A. & Kaul, S. Farnesyl pyrophosphate synthase: a key enzyme in isoprenoid biosynthetic pathway and potential molecular target for drug development. *N Biotechnol* **30,** 114-23 (2013).
47. Lange, B.M., Rujan, T., Martin, W. & Croteau, R. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc Natl Acad Sci U S A* **97,** 13172-7 (2000).
48. Odom, A.R. Five questions about non-mevalonate isoprenoid biosynthesis. *PLoS Pathog* **7,** e1002323 (2011).

49. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C. & Wishart, D.S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **39,** D1035-41 (2011).
50. Kimmig, J. [Gerhard Domagk, 1895-1964. Contribution to the chemotherapy of bacterial infections]. *Internist (Berl)* **10,** 116-20 (1969).
51. Pollak, N., Dolle, C. & Ziegler, M. The power to reduce: pyridine nucleotides--small molecules with a multitude of functions. *Biochem J* **402,** 205-18 (2007).
52. Sauve, A.A. NAD+ and Vitamin B3: From Metabolism to Therapies. *Journal of Pharmacology and Experimental Therapeutics* **324,** 883-893 (2008).
53. Khedkar, S.A., Malde, A.K. & Coutinho, E.C. Comparative protein modeling of methionine S-adenosyltransferase (MAT) enzyme from Mycobacterium tuberculosis: a potential target for antituberculosis drug discovery. *J Mol Graph Model* **23,** 355-66 (2005).
54. Perez-Leal, O., Moncada, C., Clarkson, A.B. & Merali, S. Pneumocystis S-adenosylmethionine transport: a potential drug target. *Am J Respir Cell Mol Biol* **45,** 1142-6 (2011).
55. Peleg, A.Y. & Hooper, D.C. Hospital-acquired infections due to Gram-negative bacteria. *N Engl J Med* **362,** 1804-13 (2010).
56. Chopra, I., Schofield, C., Everett, M., O'Neill, A., Miller, K., Wilcox, M., Frere, J.M., Dawson, M., Czaplewski, L., Urleb, U. & Courvalin, P. Treatment of health-care-associated infections caused by Gram-negative bacteria: a consensus statement. *Lancet Infect Dis* **8,** 133-9 (2008).
57. Wang, J., Soisson, S.M., Young, K., Shoop, W., Kodali, S., Galgoci, A., Painter, R., Parthasarathy, G., Tang, Y.S., Cummings, R., Ha, S., Dorso, K., Motyl, M., Jayasuriya, H., Ondeyka, J., Herath, K., Zhang, C., Hernandez, L., Allocco, J., Basilio, A., Tormo, J.R., Genilloud, O., Vicente, F., Pelaez, F., Colwell, L., Lee, S.H., Michael, B., Felcetto, T., Gill, C., Silver, L.L., Hermes, J.D., Bartizal, K., Barrett, J., Schmatz, D., Becker, J.W., Cully, D. & Singh, S.B. Platensimycin is a selective FabF inhibitor with potent antibiotic properties. *Nature* **441,** 358-61 (2006).
58. Brinster, S., Lamberet, G., Staels, B., Trieu-Cuot, P., Gruss, A. & Poyart, C. Type II fatty acid synthesis is not a suitable antibiotic target for Gram-positive pathogens. *Nature* **458,** 83-86 (2009).
59. Parsons, J.B., Broussard, T.C., Bose, J.L., Rosch, J.W., Jackson, P., Subramanian, C. & Rock, C.O. Identification of a two-component fatty acid kinase responsible for host fatty acid incorporation by Staphylococcus aureus. *Proceedings of the National Academy of Sciences* **111,** 10532-10537 (2014).
60. Heath, R.J. & Rock, C.O. Fatty acid biosynthesis as a target for novel antibacterials. *Curr Opin Investig Drugs* **5,** 146-53 (2004).
61. Spalding, M.D. & Prigge, S.T. Lipoic acid metabolism in microbial pathogens. *Microbiol Mol Biol Rev* **74,** 200-28 (2010).
62. Cuthbert, J.A. & Lipsky, P.E. Inhibition by 6-fluoromevalonate demonstrates that mevalonate or one of the mevalonate phosphates is necessary for lymphocyte proliferation. *Journal of Biological Chemistry* **265,** 18568-18575 (1990).
63. Wilding, E.I., Brown, J.R., Bryant, A.P., Chalker, A.F., Holmes, D.J., Ingraham, K.A., Iordanescu, S., So, C.Y., Rosenberg, M. & Gwynn, M.N. Identification, Evolution, and Essentiality of the Mevalonate Pathway for Isopentenyl Diphosphate Biosynthesis in Gram-Positive Cocci. *Journal of Bacteriology* **182,** 4319-4327 (2000).

Chapter 4

Post-vaccine microevolution of invasive *Streptococcus pneumoniae*

Amelieke J.H. Cremers*
Fredrick M. Mobegi*
Marien I. de Jonge
Sacha A.F.T. van Hijum
Jacques F. Meis
Peter W.M. Hermans
Gerben Ferwerda
Stephen D. Bentley
Aldert L. Zomer

*authors contributed equally

Scientific Reports 2015; 5:14952.

Abstract

The 7-valent pneumococcal conjugated vaccine (PCV7) has affected the genetic population of *Streptococcus pneumoniae* in pediatric carriage. Little is known however about pneumococcal population genomics in adult invasive pneumococcal disease (IPD) under vaccine pressure. We sequenced and serotyped 349 strains of *S. pneumoniae* isolated from IPD patients in Nijmegen between 2001 and 2011. Introduction of PCV7 in the Dutch National Immunization Program in 2006 preluded substantial alterations in the IPD population structure caused by serotype replacement. No evidence could be found for vaccine induced capsular switches. We observed that after a temporary bottleneck in gene diversity after the introduction of PCV7, the accessory gene pool re-expanded mainly by genes already circulating pre-PCV7. In the post-vaccine genomic population a number of genes changed frequency, certain genes became overrepresented in vaccine serotypes, while others shifted towards non-vaccine serotypes. Whether these dynamics in the invasive pneumococcal population have truly contributed to invasiveness and manifestations of disease remains to be further elucidated. We suggest the use of whole genome sequencing for surveillance of pneumococcal population dynamics that could give a prospect on the course of disease, facilitating effective prevention and management of IPD.

Introduction

The bacterial pathogen *Streptococcus pneumoniae* remains an important cause of pneumonia and meningitis worldwide, causing an estimated 1.6 million deaths annually [1] Immunization with a 7-valent pneumococcal conjugated vaccine (PCV), both in young children [2] and the elderly [3], has been demonstrated to be highly efficacious in preventing invasive pneumococcal disease (IPD) caused by vaccine serotypes. Moreover, after the introduction of routine pediatric vaccination, an additional decrease in vaccine-type pneumococcal infections occurred, both in pediatric carriage [4, 5] as well as in IPD at all ages [6], indicative of a herd immunity effect. However, in all regions where routine PCV vaccination has been introduced, we see replacement in pneumococcal carriage and disease by non-vaccine serotypes [7-9], which at least partially abrogates the preventive effect of vaccination. In the Netherlands, the number of reported cases of IPD has not decreased after vaccination due to full replacement by non-vaccine serotypes [10]

Currently available pneumococcal vaccines target the pneumococcal polysaccharide capsule. Although over 90 antigenically distinct pneumococcal capsules (denominated as serotypes) are known to date, only a selection of 7 serotypes (later extended to 10 and 13) is included in the PCV, based on modelling vaccination effects by the frequency and the propensity for invasiveness for each serotype [11, 12]. Whereas the polysaccharide capsule is essential for the pneumococcus to cause IPD [13], recent publications suggest that true determinants of invasiveness may lay underneath the pneumococcal capsule, represented by variations on the pneumococcal genome [14-18], which should be included in modelling vaccine effects [19].

Since the introduction of routine pediatric immunization with PCV, several studies have detailed pneumococcal genomic epidemiology in pediatric carriage and otitis media under vaccine pressure. They all demonstrated that the pneumococcal genomic structure of pediatric carriage remained fairly stable, and that serotype replacement occurred mainly through expansion of previously existent clones [20-25]. The single study that examined whole pneumococcal genomes, including virulence factors, reported little effect on the accessory genome at the overall pneumococcal carriage population level despite massive serotype replacement [26]. However, the effect of pneumococcal vaccination on whole genome epidemiology of invasive pneumococcal disease remains unexplored, although it may hold invaluable information on understanding the long term effects of mass vaccination, especially with regard to changes in clinical manifestation of disease due to the changing prevalence of virulence factors in the pneumococcal population.

PCV7 was introduced in the Dutch National Immunization Program in April 2006 while PCV10 was introduced in May 2011. In this study, we have sequenced 349 invasive pneumococcal isolates from 2001 to June 2011, in order to investigate the herd-immunity effect of PCV7 on pneumococcal population genomics in IPD.

Results

Cohort and serotype dynamics

We serotyped and sequenced 349 strains of *S. pneumoniae* isolated from an unbiased cohort of IPD patients in Nijmegen, the Netherlands, between 2001 and 2011. In the Netherlands, the elderly are not routinely vaccinated against pneumococcal disease, pediatric pneumococcal vaccination coverage has been high (>94%) [27], and penicillin resistance was below 3% during the entire study period [28]. These facts increase the probability that in this cohort, major changes in adult IPD epidemiology are due to herd immunity after pediatric immunization. Moreover, risk factors for acquiring a pneumococcal infection have not changed for the IPD patients in our study cohort, suggesting that replacement has been a merely pneumococcus-mediated phenomenon. According to our analysis, since the 7-valent PCV (PCV7) was introduced in the pediatric Dutch National Immunization Program in April 2006, the invasive pneumococcal population structure has altered substantially. Its introduction preluded a steady decrease in adult IPD cases caused by the seven serotypes it protects against (vaccine types, VTs): 4, 6B, 9V, 14, 18C, 19F, and 23F, supporting a herd immunity effect of PCV7, concordant with a previous nationwide study [29]. In addition, the absolute increase in IPD cases caused by non-PCV7 protected serotypes (non-vaccine types, nVTs) observed in the Netherlands [30], as well as within our cohort [31] is suggestive of serotype replacement or capsular switching. Similarly to the pediatric pneumococcal carriage study by Croucher *et al* [26], in our non-vaccinated IPD cohort we observed an increase of subtype variants of VTs: 6A, 19A, 23A, and 23B. However, the post PCV7 nVT IPD cases were dominated by serotypes 1, 3, 7F, and 8 which may be seen less frequently in carriage cohorts [11].

Core genome phylogeny

Phylogeny of the 349 IPD isolates was reconstructed based on a 'superalignment' of concatenated alignments of genes in 786 out of the 1075 core orthologous groups (OGs) containing genes present in a single copy in all isolates. 289 OGs with a dissimilar tree topology and branch distance as compared to ribosomal protein encoding genes were excluded as their phylogeny may be influenced by recombination or homoplasies. A tight clustering of isolates by serotype was observed (Figure 1). Similar to the carriage study by Croucher *et al* [26], in our IPD study, serotype replacement was responsible for the increase in nVTs. The rare cases of capsular switch were probably extant before immunization and did not increase upon the introduction of PCV7. It is therefore unlikely that capsular switching was induced by vaccine pressure, although larger numbers of isolates may be needed to confirm this. Compared to their closest neighbors in the phylogenetic tree, capsular switched strains accumulated only between 1 and 10 single nucleotide polymorphisms (SNPs) in their core genome.

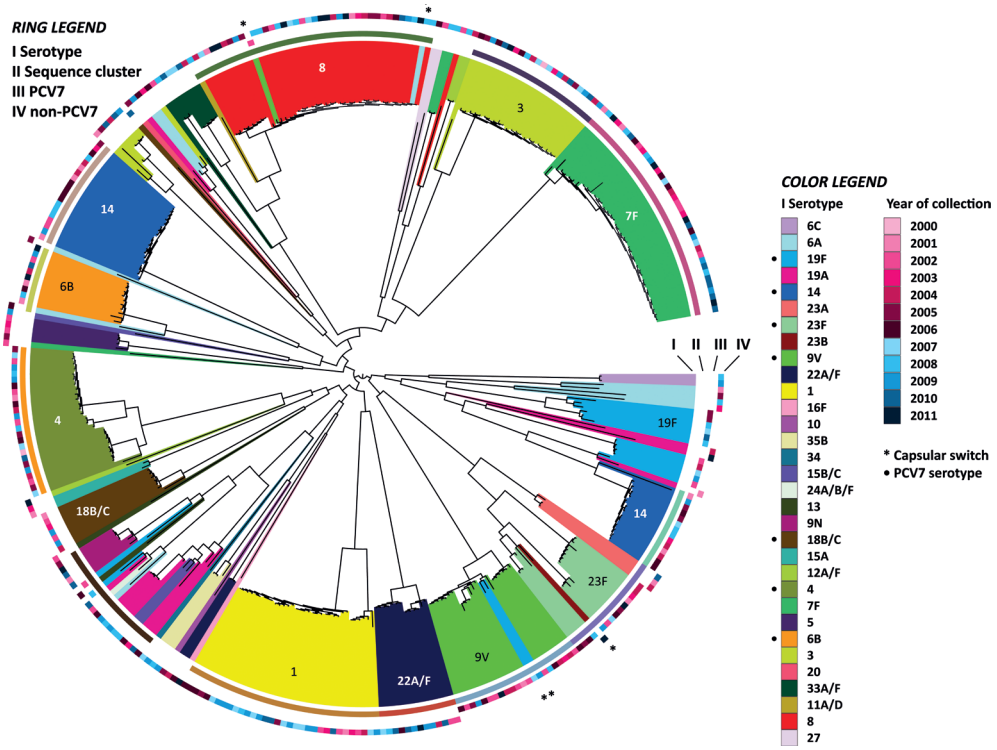


Figure 1. Structure of the invasive pneumococcal population. The maximum likelihood phylogeny was generated using 59,682 polymorphic sites within a 970,559 bp codon alignment of 755 core OGs with similar phylogenetic profiles as ribosomal protein encoding genes in order to exclude genes that were acquired by, for example, horizontal gene transfer. The inner bars each represent a pneumococcal strain (I) and are colored by capsular serotype. The second circle displays the sequence clusters (II). The year of collection is marked by the color (red: pre-PCV7, blue: post-PCV7) and intensity of coloration per year, for PCV7 vaccine serotypes (III) and non-vaccine serotypes (IV).

The majority of these SNPs occurred in regions centered on genes involved in resistance to and efflux of antibiotics, synthesis of cell wall proteins, as well as transposases, pili and phage elements (Supplementary Table 1). A similar situation with regard to gene insertions or deletions was observed in the capsule switched isolates, suggesting phages and transposing elements play a major role in shaping the pneumococcal genome on relatively short timescales (Supplementary Table 2). The sequence clusters were based on BAPS analyses, and were determined to study if serotype replacement within a serogroup would involve closely related core genomes. Although this holds true for serogroup 23, replacement in serogroups 6 and 19 is scattered over the phylogenetic tree.

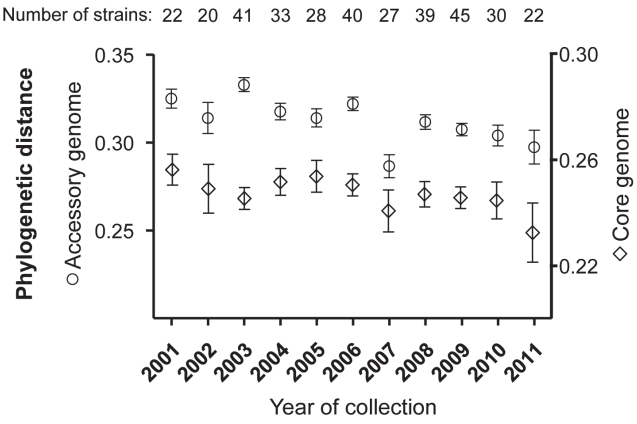


Figure 2. Annual diversity among accessory genomes. The diversity was calculated between strains collected within the same year. Using a binary measure of the presence (1) or absence (0) of a gene in each accessory OG, diversity among strains was calculated by applying the distance measure described by Dutilh *et al* [50]. Core OG diversity was determined by comparing the pairwise distances between isolates from the phylogenetic tree per year. Circles: denote accessory distance; diamonds: denote core OG distance; whiskers: 95% confidence intervals.

Genome diversity

Analysis of the diversity of the accessory genome among strains revealed a significant drop in OG diversity in 2007 shortly after the introduction of PCV7 ($p < 0.0001$), and a re-expansion in subsequent years (Figure. 2); a similar trend is seen for the core OG diversity, however here the 95% CI overlap. This illustrates that vaccination had an early temporary bottleneck effect on IPD gene diversity. In concordance with the decreased diversity after vaccination, a considerable number of OGs were significantly altered in prevalence, but most of them returned towards equilibrium afterwards (Figure. 3). Majority of the OGs that contributed to the re-expansion of diversity in the accessory genome after 2007 was comparable to those circulating pre PCV7. The decreased diversity after vaccination when VT stains were steadily removed from the population seems to represent the simultaneous decrease of certain sets of VT-related genes. However, the re-expansion of diversity with similar genes, suggests that those might now be carried by replacing nVT strains. Such a phenomenon has recently been predicted by modeling from pneumococcal carriage genome data [32]. Aside from this tendency to recover, certain OGs dispersed from their prevalence in the original gene pool. Both dynamics will be detailed below. A second gradual decrease in accessory genome diversity was observed towards 2011. This was not due to a distinct decreasing diversity among VT strains post vaccination, as the diversity in the accessory genomes remained similar between VT and nVT strains (see Supplementary Figure 1).

Gene dynamics

For the entire cohort, 506 pneumococcal genes were significantly overrepresented in the group of VTs, and 223 in nVTs (see Supplementary Table 3). This observation suggests that the distribution of particular genes over strains is more clustered in the VT group. Given the low number of serotypes in the VT group, it may indicate that also the phylogeny of the accessory genome clusters by serotype. Whereas 97 genes were

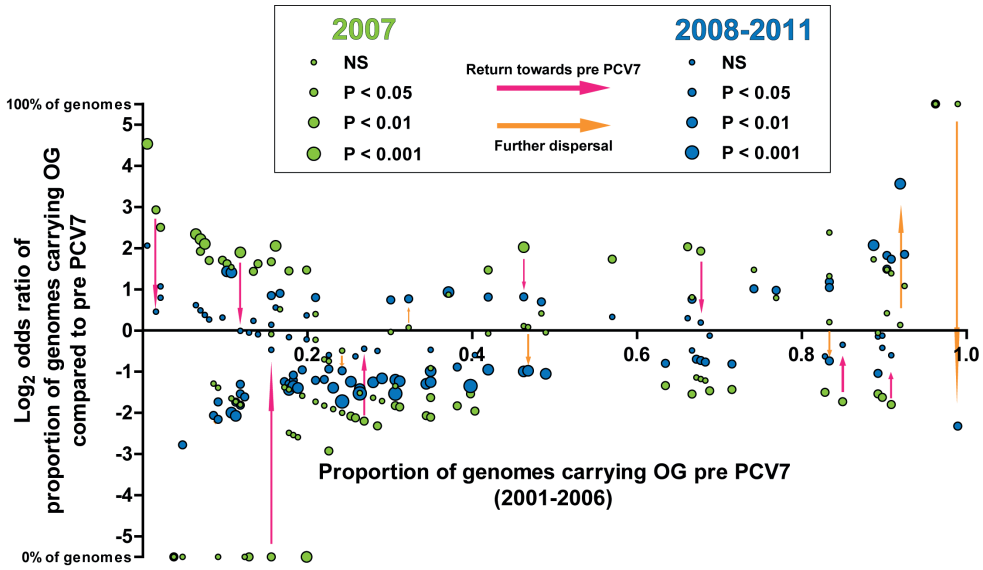


Figure 3. Temporal stability of accessory gene pool. Odds ratios for prevalence of individual OGs in the accessory genome whose frequency significantly altered at bottleneck shortly post PCV7 (2007) or during re-expansion (2008-2011), compared to pre PCV7 (2001-2006) The X-axis indicates the proportion of isolates carrying each OG in 2001 while the Y-axis indicates the log transformed odds ratio of strains carrying each OG in 2007 (green dots), and post PCV7 (blue dots), relative to pre PCV7.

exclusively found in the VTs and 331 genes were unique to the nVTs, the average number of unique genes per serotype was equal.

Post PCV7, 105 genes significantly decreased in frequency, among which many phage-related genes and transposons that went down along with VTs carrying them. Only 33 genes expanded significantly, including mainly ABC-transporters, zinc metallopeptidases and genes facilitating recombination. The fact that VT strains had more OGs associated with phages throughout the course of the study may relate to the 'kill the winner' strategy employed by phages as VT strains had a higher prevalence and would therefore be more likely to be a target for phage predation.

Under the selective pressure of herd immunity post PCV7, the rate of adult IPD caused by VTs decreased. Those VT strains that still caused IPD after the PCV7 bottleneck more often carried OG0185 DNA methylase (1.8 fold increase, p=0.008), OG1539 flippase wzx associated with serogroup 6 (2.3 fold increase, p=0.037), and a cluster consisting of 13 Tn5253-related genes (3.7 fold increase, p=0.02). Although the latter are often related to acquisition of antibiotic resistance, all strains concerned were fully susceptible to tetracycline and clindamycin, and only 1 was resistant against erythromycin and clarithromycin. The patients involved had no notable history indicative of increased

antibiotic pressure (antibiotics use before admission, comorbidities associated with frequent antibiotics use, nursing home residency). Therefore, presence of the Tn5253 cluster may have facilitated maintenance in IPD through a different mechanism.

After 2007 the number of adult IPD cases caused by non-vaccine serotypes expanded. Among nVT IPD, the proportion of strains carrying OG1150 (toxin antitoxin component HicA) increased post PCV7 ($p=0.032$). Genes that were first observed in nVTs after PCV7 ($n=115$) were infrequent. They coded for phages, transposons and membrane structures, but also included OG2585 (endonuclease), and OG2354 (multidrug resistance transporter MdtK). Although the latter is homologous to the MepA multidrug transporter of *S. aureus*, here a different antibiotic resistance pattern was observed. Of the two strains carrying MdtK, one isolate was fully susceptible *in vitro*, while the other was resistant to clarithromycin, and intermediately susceptible to doxycycline and ofloxacin, but not resistant to fluoroquinolones - the substrate of MepA in *S. aureus* [31].

Although capsular switching has been rare, theoretically, individual pneumococcal genes may also benefit from translocation to strains with a polysaccharide capsule not targeted by vaccination. We studied whether individual genes shifted from VT to nVT strains by comparing the odds ratios for linkage to VT versus nVT strains before and after the PCV7 bottleneck. Of the 86 genes that displayed increased odds for being related to nVTs after PCV7, 72% were previously related to VTs, indicative of a certain originality to nVT strains. Whether these genes actually benefit from a transition towards nVT strains, or contribute to severity of infection remains to be elucidated.

Although after the introduction of PCV7 the overall severity of IPD cases (measured by pneumonia severity index) remained stable, higher rates of pleural effusion and empyema have been reported [31, 34], which may be attributed to a changed frequency of virulence factors in the pneumococcal population. For instance zinc metalloproteinase C (*zmpC*), mainly present in nVTs, has been associated with a more severe clinical manifestation of IPD and was shown to be expanding [17]. This illustrates that elimination of VTs may result in substitution by nVT strains that potentially confer more severe disease. Modeling of vaccination-induced changes in frequency of disease associated genes may therefore be of interest to optimize target selection in new generations of pneumococcal vaccines. Such an approach would require further exploration of the role of pneumococcal genotypes in the manifestation of human infection.

Conclusions

Despite serotype replacement in pneumococcal disease, after pediatric pneumococcal vaccination with PCV7 we observed a temporary bottleneck in gene diversity, which re-expanded mainly by genes already present in the original gene pool. Our observations suggest that the introduction of PCV7 has only temporarily affected the pneumococcal population, like disease frequency, with the effects lasting up to one year. Certain genes

in IPD have dispersed from their original prevalence, while others became overrepresented in VT, or shifted towards nVT. These changes may influence clinical manifestation of disease if these genes are associated with disease. We suggest whole genome sequencing to investigate pneumococcal dynamics after vaccination and as such maintain close surveillance of strains in a population; information that could give a prospect on an altered course and severity of disease, facilitating effective prevention and management of invasive pneumococcal disease.

Methods

Pneumococcal strain collection, DNA isolation and serotyping

Pneumococcal strains were isolated from adults hospitalized with a bacteremic pneumococcal infection at two Dutch hospitals between January 2000 and June 2011 as described by Cremers *et al* [31]. The candidates were retrospectively included in the Pneumococcal Bacteremia Collection Nijmegen (PBCN). This observational cohort study was approved by the Local Medical Ethics Committees of both participating hospitals. The cohort was divided in pre and post PCV-7 strains as defined by their collection from blood before or after January 1 2007. The pneumococcal serotypes were determined using the multiplex PCR analysis [35]. In case multiplex PCR was inconclusive, serotyping was complemented with the Quellung reaction using Pneumococcus Neufeld Antisera (Statens Serum Institute, Copenhagen, Denmark) according to the manufacturer's instructions [31]. All pneumococcal serotypes were then confirmed using an in-house implementation of the Sanger Institute molecular capsular typing (MCT) system [36, 37].

Genomic DNA preparation and whole genome sequencing

The strains were grown statically in 10ml of Todd Hewitt broth (Merck, Darmstadt, Germany) with 5% yeast extract at 37°C and 5% CO_2 to an optical density at 620nm of 0.2-0.3. The bacterial pellet was washed with 1ml PBS and DNA was extracted using QIAGEN Genomic-tips 20/G and Genomic DNA Buffer Set (both Qiagen, Venlo, The Netherlands) according to the manufacturer's instructions for mini DNA preparations. The concentration of extracted genomic DNA was determined using Quant-iT™ PicoGreen® dsDNA Reagent (LifeTechnologies, Bleiswijk, The Netherlands) and the TECAN GENios plate reader with Magellan software (Tecan, Giessen, The Netherlands) and its intactness was confirmed with gel electrophoresis. Genomic DNA was sequenced on an Illumina HiSeq 2000 as paired-end reads of 100 nucleotides.

Genome assembly, mapping and annotation

The strains were assembled using the Sanger Institute genomes assembly pipeline. For each strain, Velvet [38] was used to create multiple assemblies by varying the kmer size

between 66% and 90% of the read length. From these assemblies, the one with the highest N50 was chosen and contigs that were shorter than the insert size length were removed. The resulting assembly was improved by the following steps: The contigs of the assembly were scaffolded by iteratively running SSPACE using default settings [39]. Then, gaps identified as 1 or more N's, were targeted for closure by running 120 iterations of GapFiller [40]. Genomes were annotated using an in-house implementation of Prokka [41]. The genome assemblies were deposited at the European Nucleotide Archive under study number ERP001789.

Gene clustering and phylogenetic analysis

All putative coding sequences were translated and analyzed by an all-versus-all blastP employing a 10E-15 *e-value* cut-off and BLOSUM90 matrix. TribeMCL [42] was used to cluster orthologs by an inflation value of 4 to implement the MCL step, which resulted into a total of 3021 OGs. Of these OGs, 1075 were denominated as core OGs as they consisted of proteins present in single copies in each of the 349 strains. Protein sequences of these core OGs were aligned with MUSCLE [43], and subsequently codon translated into nucleotide alignments [44]. Genes of OGs encoding ribosomal proteins from each strain were concatenated into a single 'ribosomal' alignment. A reference maximum likelihood phylogeny was generated based on this ribosomal alignment using RAxML [45]. Phylogenies for each of the remaining core OGs were also generated separately. Plotting the Euclidian distances (EuD) of each OG tree with the ribosomal protein encoding tree resulted in three distributions of distances. To reach an alignment with high phylogenetic resolution, the sequences from OGs whose phylogenies were similar to the ribosomal phylogeny (EuD ≤ 0.03) were concatenated to the ribosomal alignment to give a single reduced 'core' super-alignment. Finally, polymorphic regions of this super-alignment were extracted and re-analyzed with RAxML using a general time-reversible (GTR) nucleotide substitution model. Core OG diversity was determined by comparing the average of the pairwise distances between the isolates per year. The reduced core super-alignment was also analyzed using BAPS; Bayesian Analysis of Population Structure software [46] to determine the sequence clusters. Two runs of 40 and 50 maximum clusters were performed, each creating 12 largely monophyletic sequence clusters and an extra 'sink' cluster that incorporated all unclassified isolates. Visualization of the tree was performed using iTOL [47].

Analysis of the population structure

Core genomes and whole genomes of strains with a capsule switched serotype were separately aligned in Mauve [48], along with their closest neighbors (Supplementary File 1). SNP coordinates were isolated and characterized in Artemis [49].

Genomic diversity and dynamics

The genomic diversity was determined based on the strains' accessory genomes. For each OG a binary measure of presence (1) or absence (0) of a representative protein from a given strain was generated. The contribution of each strain to an OG was therefore denoted by a single numeric value (1 or 0) to represent the presence or absence of a gene, or a group of paralogs in a certain strain. This information was collated into a gene presence/absence matrix (Supplementary File 2). Using the method described by Dutilh *et al* [50], the genomic variations among strains collected within each year were calculated. Core OGs (those constituted of a single gene from each of the analyzed isolates) were excluded from this analysis.

The time frames applied to detect changes in the proportion of strains carrying a certain OG post PCV7 compared to before were 2001 to 2006; pre-vaccine, and 2008 to 2011; post-vaccine respectively.

Statistical analysis

The difference in phylogenetic distance among strains collected between 2001 to 2006 and 2007 was tested by an unpaired t-test. Differences in pre- and post-vaccine gene frequencies were tested by Fisher Exact test. For all analyses, the significance level was set at 0.05.

Acknowledgements

This work was supported by the Dutch Government (AgentschapNL) and the European Union Seventh Framework Programme (ENIAC Joint Undertaking - CAJAL4EU project). We sincerely thank the staff of the Canisius-Wilhelmina Hospital in Nijmegen, for facilitating the collection of pneumococcal isolates and clinical data. We are also grateful to Dr. Arie van der Ende at The Netherlands Reference Laboratory for Bacterial Meningitis for providing specific pneumococcal strains. Finally, we acknowledge Feyruz Yalcin at the Sanger Institute in Cambridge and Bas Dutilh at the University of Utrecht, for their assistance in data management and phylogenetic diversity analysis respectively.

Supporting data

Supplementary material for this chapter are available in the online publication at: http://www.nature.com/articles/srep14952#supplementary-information

References

1. World Health Organization. Pneumococcal vaccines. *Weekly epidemiological record* **78**, 110-119 (2003).
2. Pavia, M., Bianco, A., Nobile, C.G., Marinelli, P. & Angelillo, I.F. Efficacy of pneumococcal vaccination in children younger than 24 months: a meta-analysis. *Pediatrics* **123**, e1103-10 (2009).
3. Bonten, M.J.M., Huijts, S.M., Bolkenbaas, M., Webber, C., Patterson, S., Gault, S., van Werkhoven, C.H., van Deursen, A.M.M., Sanders, E.A.M., Verheij, T.J.M., Patton, M., McDonough, A., Moradoghli-Haftvani, A., Smith, H., Mellelieu, T., Pride, M.W., Crowther, G., Schmoele-Thoma, B., Scott, D.A., Jansen, K.U., Lobatto, R., Oosterman, B., Visser, N., Caspers, E., Smorenburg, A., Emini, E.A., Gruber, W.C. & Grobbee, D.E. Polysaccharide Conjugate Vaccine against Pneumococcal Pneumonia in Adults. *New England Journal of Medicine* **372**, 1114-1125 (2015).
4. Dagan, R., Givon-Lavi, N., Zamir, O., Sikuler-Cohen, M., Guy, L., Janco, J., Yagupsky, P. & Fraser, D. Reduction of nasopharyngeal carriage of *Streptococcus pneumoniae* after administration of a 9-valent pneumococcal conjugate vaccine to toddlers attending day care centers. *J Infect Dis* **185**, 927-36 (2002).
5. Spijkerman, J., van Gils, E.J., Veenhoven, R.H., Hak, E., Yzerman, E.P., van der Ende, A., Wijmenga-Monsuur, A.J., van den Dobbelsteen, G.P. & Sanders, E.A. Carriage of *Streptococcus pneumoniae* 3 years after start of vaccination program, the Netherlands. *Emerg Infect Dis* **17**, 584-91 (2011).
6. Lexau, C.A., Lynfield, R., Danila, R., Pilishvili, T., Facklam, R., Farley, M.M., Harrison, L.H., Schaffner, W., Reingold, A., Bennett, N.M., Hadler, J., Cieslak, P.R. & Whitney, C.G. Changing epidemiology of invasive pneumococcal disease among older adults in the era of pediatric pneumococcal conjugate vaccine. *JAMA* **294**, 2043-51 (2005).
7. Miller, E., Andrews, N.J., Waight, P.A., Slack, M.P. & George, R.C. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: an observational cohort study. *Lancet Infect Dis* **11**, 760-8 (2011).
8. Weinberger, D.M., Malley, R. & Lipsitch, M. Serotype replacement in disease after pneumococcal vaccination. *Lancet* **378**, 1962-73 (2011).
9. Scott, J.R., Millar, E.V., Lipsitch, M., Moulton, L.H., Weatherholtz, R., Perilla, M.J., Jackson, D.M., Beall, B., Craig, M.J., Reid, R., Santosham, M. & O'Brien, K.L. Impact of more than a decade of pneumococcal conjugate vaccine use on carriage and invasive potential in Native American communities. *J Infect Dis* **205**, 280-8 (2012).
10. Netherlands Reference Laboratory for Bacterial Meningitis (AMC/RIVM). Bacterial meningitis in the Netherlands; annual report 2013. (2014).
11. Brueggemann, A.B., Griffiths, D.T., Meats, E., Peto, T., Crook, D.W. & Spratt, B.G. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis* **187**, 1424-32 (2003).
12. Sandgren, A., Sjostrom, K., Olsson-Liljequist, B., Christensson, B., Samuelsson, A., Kronvall, G. & Henriques Normark, B. Effect of clonal and serotype-specific properties on the invasive capacity of *Streptococcus pneumoniae*. *J Infect Dis* **189**, 785-96 (2004).
13. Watson, D.A. & Musher, D.M. Interruption of capsule production in Streptococcus pneumonia serotype 3 by insertion of transposon Tn916. *Infect Immun* **58**, 3135-8 (1990).

14. Hanage, W.P., Kaijalainen, T.H., Syrjanen, R.K., Auranen, K., Leinonen, M., Makela, P.H. & Spratt, B.G. Invasiveness of serotypes and clones of *Streptococcus pneumoniae* among children in Finland. *Infect Immun* **73,** 431-5 (2005).
15. Blomberg, C., Dagerhamn, J., Dahlberg, S., Browall, S., Fernebro, J., Albiger, B., Morfeldt, E., Normark, S. & Henriques-Normark, B. Pattern of accessory regions and invasive disease potential in *Streptococcus pneumoniae*. *J Infect Dis* **199,** 1032-42 (2009).
16. Browall, S., Norman, M., Tangrot, J., Galanis, I., Sjostrom, K., Dagerhamn, J., Hellberg, C., Pathak, A., Spadafina, T., Sandgren, A., Battig, P., Franzen, O., Andersson, B., Ortqvist, A., Normark, S. & Henriques-Normark, B. Intraclonal variations among *Streptococcus pneumoniae* isolates influence the likelihood of invasive disease in children. *J Infect Dis* **209,** 377-88 (2014).
17. Cremers, A.J., Kokmeijer, I., Groh, L., de Jonge, M.I. & Ferwerda, G. The role of ZmpC in the clinical manifestation of invasive pneumococcal disease. *Int J Med Microbiol* **304,** 984-9 (2014).
18. Croucher, N.J., Mitchell, A.M., Gould, K.A., Inverarity, D., Barquist, L., Feltwell, T., Fookes, M.C., Harris, S.R., Dordel, J., Salter, S.J., Browall, S., Zemlickova, H., Parkhill, J., Normark, S., Henriques-Normark, B., Hinds, J., Mitchell, T.J. & Bentley, S.D. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. *PLoS Genet* **9,** e1003868 (2013).
19. Klugman, K.P., Bentley, S.D. & McGee, L. Determinants of invasiveness beneath the capsule of the pneumococcus. *J Infect Dis* **209,** 321-2 (2014).
20. Lipsitch, M., O'Neill, K., Cordy, D., Bugalter, B., Trzcinski, K., Thompson, C.M., Goldstein, R., Pelton, S., Huot, H., Bouchet, V., Reid, R., Santosham, M. & O'Brien, K.L. Strain characteristics of *Streptococcus pneumoniae* carriage and invasive disease isolates during a cluster-randomized clinical trial of the 7-valent pneumococcal conjugate vaccine. *J Infect Dis* **196,** 1221-7 (2007).
21. Bogaert, D., Veenhoven, R.H., Sluijter, M., Wannet, W.J., Rijkers, G.T., Mitchell, T.J., Clarke, S.C., Goessens, W.H., Schilder, A.G., Sanders, E.A., de Groot, R. & Hermans, P.W. Molecular epidemiology of pneumococcal colonization in response to pneumococcal conjugate vaccination in children with recurrent acute otitis media. *J Clin Microbiol* **43,** 74-83 (2005).
22. Gherardi, G., D'Ambrosio, F., Visaggio, D., Dicuonzo, G., Del Grosso, M. & Pantosti, A. Serotype and clonal evolution of penicillin-nonsusceptible invasive *Streptococcus pneumoniae* in the 7-valent pneumococcal conjugate vaccine era in Italy. *Antimicrob Agents Chemother* **56,** 4965-8 (2012).
23. Hanage, W.P., Bishop, C.J., Huang, S.S., Stevenson, A.E., Pelton, S.I., Lipsitch, M. & Finkelstein, J.A. Carried pneumococci in Massachusetts children: the contribution of clonal expansion and serotype switching. *Pediatr Infect Dis J* **30,** 302-8 (2011).
24. Gladstone, R.A., Jefferies, J.M., Tocheva, A.S., Beard, K.R., Garley, D., Chong, W.W., Bentley, S.D., Faust, S.N. & Clarke, S.C. Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine* **33,** 2015-21 (2015).
25. Chang, Q., Stevenson, A.E., Croucher, N.J., Lee, G.M., Pelton, S.I., Lipsitch, M., Finkelstein, J.A. & Hanage, W.P. Stability of the pneumococcal population structure in Massachusetts as PCV13 was introduced. *BMC Infect Dis* **15,** 68 (2015).
26. Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P. & Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45,** 656-63 (2013).

27. van Lier, E., Oomen, P., Conyn-van Spaendonck, M., Drijfhout, I., Zonnenberg-Hoff, I. & de Melker, H. Immunisation coverage National Immunisation Programme in the Netherlands. *National Institute for Public Health and the Environment* (2014).
28. The Dutch Working Party on Antibiotic Policy. Monitoring of Antimicrobial Resistance and Antibiotic Usage in Animals in the Netherlands in 2012. www.swab.nl (2013).
29. Elberse, K.E., van der Heide, H.G., Witteveen, S., van de Pol, I., Schot, C.S., van der Ende, A., Berbers, G.A. & Schouls, L.M. Changes in the composition of the pneumococcal population and in IPD incidence in The Netherlands after the implementation of the 7-valent pneumococcal conjugate vaccine. *Vaccine* **30,** 7644-51 (2012).
30. van Deursen, A.M., van Mens, S.P., Sanders, E.A., Vlaminckx, B.J., de Melker, H.E., Schouls, L.M., de Greeff, S.C. & van der Ende, A. Invasive pneumococcal disease and 7-valent pneumococcal conjugate vaccine, the Netherlands. *Emerg Infect Dis* **18,** 1729-37 (2012).
31. Cremers, A.J., Meis, J.F., Walraven, G., Jongh, C.E., Ferwerda, G. & Hermans, P.W. Effects of 7-valent pneumococcal conjugate 1 vaccine on the severity of adult 2 bacteremic pneumococcal pneumonia. *Vaccine* **32,** 3989-94 (2014).
32. Watkins, E.R., Penman, B.S., Lourenco, J., Buckee, C.O., Maiden, M.C. & Gupta, S. Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *PLoS Pathog* **11,** e1005034 (2015).
33. Kaatz, G.W., McAleese, F. & Seo, S.M. Multidrug resistance in *Staphylococcus aureus* due to overexpression of a novel multidrug and toxin extrusion (MATE) transport protein. *Antimicrob Agents Chemother* **49,** 1857-64 (2005).
34. Burgos, J., Lujan, M., Falco, V., Sanchez, A., Puig, M., Borrego, A., Fontanals, D., Planes, A.M., Pahissa, A. & Rello, J. The spectrum of pneumococcal empyema in adults in the early 21st century. *Clin Infect Dis* **53,** 254-61 (2011).
35. Pai, R., Gertz, R.E. & Beall, B. Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates. *J Clin Microbiol* **44,** 124-31 (2006).
36. Kong, F., Wang, W., Tao, J., Wang, L., Wang, Q., Sabananthan, A. & Gilbert, G.L. A molecular-capsular-type prediction system for 90 *Streptococcus pneumoniae* serotypes using partial cpsA-cpsB sequencing and wzy- or wzx-specific PCR. *J Med Microbiol* **54,** 351-6 (2005).
37. Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrell, B., Reeves, P.R., Parkhill, J. & Spratt, B.G. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2,** e31 (2006).
38. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18,** 821-9 (2008).
39. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578-9 (2011).
40. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biology* **13,** R56 (2012).
41. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068-9 (2014).
42. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30,** 1575-84 (2002).

43. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32,** 1792-7 (2004).
44. Wernersson, R. & Pedersen, A.G. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* **31,** 3537-9 (2003).
45. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456-63 (2005).
46. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying Currents in the Gene Pool for Bacterial Populations Using an Integrative Approach. *PLoS Computational Biology* **5,** e1000455 (2009).
47. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39,** W475-8 (2011).
48. Darling, A.C., Mau, B., Blattner, F.R. & Perna, N.T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14,** 1394-403 (2004).
49. Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J. & Rajandream, M.A. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672-6 (2008).
50. Dutilh, B.E., Huynen, M.A., Bruno, W.J. & Snel, B. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol* **58,** 527-39 (2004).

Chapter 5

Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data

Fredrick M. Mobegi
Sacha A. F. T. van Hijum
Amelieke J.H Cremers
Marien I. de Jonge
Stephen D. Bentley
Aldert Zomer

Manuscript submitted for publication

Abstract

Advances in genome sequencing technologies and genome-wide association studies (GWAS) have provided unprecedented insights into the molecular basis of microbial phenotypes and enabled the identification of the underlying genetic variants in real populations. However, utilization of genome sequencing in clinical phenotyping of bacteria is challenging due to the lack of reliable and accurate approaches. Here, we report a method for predicting microbial resistance patterns using genome sequencing data. We analyzed whole genome sequences of 1,680 *Streptococcus pneumoniae* isolates from four independent populations using GWAS and identified probable hotspots of genetic variation which correlate with phenotypes of resistance to essential classes of antibiotics. With the premise that accumulation of putative resistance-conferring SNPs, potentially in combination with specific resistance genes, precedes full resistance, we retrogressively surveyed the hotspot loci and quantified the number of SNPs and/or genes, which if accumulated would confer full resistance to an otherwise susceptible strain. We name this approach the 'distance to resistance'. It can be used to identify the creep towards complete antibiotics resistance in bacteria using genome sequencing. This approach serves as a basis for the development of future sequencing-based methods for predicting resistance profiles of bacterial strains in hospital microbiology and public health settings.

Introduction

Streptococcus pneumoniae, or the pneumococcus, is part of the normal bacterial flora of the human nasopharynx, but can occasionally infiltrate sterile sites of the body progressing to disease [1]. An estimated 1.6 million deaths associated with *S. pneumoniae* are reported every year worldwide, mostly affecting children under five years [2]. Despite the immunization efforts reducing pneumococcal disease, there is merely a marginal prospect of eliminating the pneumococcal disease. This is because the available pneumococcal conjugate vaccines (PCV) only protect against a handful vaccine-type serotypes [3], and their removal facilitates rapid replacement by the non-vaccine serotypes, which consequently increase in carriage prevalence, disease, and antibiotic resistance [4, 5]. The burden of pneumococcal disease is intensified by the mounting resistance of the pneumococcus to essential antibiotics in clinical use.

Since the first case of penicillin-resistant pneumococcus was reported [6] followed by outbreaks of disease caused by multidrug-resistant pneumococci [7], the antibiotic resistance patterns of *S. pneumoniae* have drastically evolved and escalated worldwide [8]. The pneumococcus is known to be highly recombinogenic [9], allowing sequences that confer antimicrobial non-susceptibility to be readily introduced into the genome. Discovery of the genetic determinants underlying microbial phenotypes such as transmission, antimicrobial resistance, and virulence is an important question in microbiology. Traditionally, changes in DNA which are associated with antibiotic resistance in *S. pneumoniae* were identified using sequence comparison [10, 11], laboratory mutagenesis [10], and identification of horizontally transferred sequences [12]. These techniques are limited in specificity and sensitivity necessary for clinical laboratories. They only reveal common genomic regions where change has occurred in the so-called 'mosaic' genes and are narrow in their application to study actual populations and may miss out on situations where multiple mutations occurring in different genomic loci are required for full antibiotic resistance [12]. Significant advances in high-throughput genome sequencing technologies and bacterial genome-wide association studies (GWAS) now allow identification of statistical association between plausible causal genetic variants and microbial phenotypes in real populations [13]. This approach was recently used to identify the single nucleotide polymorphisms (SNPs) in DNA that may confer beta-lactam resistance in *S. pneumoniae* [14].

However, the understanding of how genetic variations contribute to antibiotics resistance remains underexplored. Here we report the use of genome sequencing and GWAS to investigate SNPs and genes associated with resistance to four essential classes of antibiotics; collectively referred to as antibiotic resistance hereafter. We name the cumulative odds ratio of these resistance-conferring variants the "distance to resistance" for the pneumococcus. We analyzed 1,680 invasive and carriage pneumococcal isolates from Nijmegen, the Netherlands [15], Massachusetts, USA [16], Maela, Thailand [14, 17], and isolates from Sickle-cell anemic children (hereafter referred to as SCD; sickle cell

disease) in the USA [18]. The genotypic and phenotypic diversity in these independent cohorts, whose draft genomes and phenotypes for antibiotic resistance are available, represent a unique paradigm for identifying probable genetic variants underlying pneumococcal antibiotic resistance. With the premise that presence of particular genes or accumulation of specific SNPs precedes full drug resistance of a fit clone, whole genome sequencing and GWAS could be used to evaluate the rate of accumulation of candidate resistance-conferring variants and provide an early warning sign of increasing antibiotic resistance. We hypothesize that the SNPs and/or resistance associated genes separating the phenotypes of antibiotic resistance are the plausible maximum number of mutations, which if accumulated could render high antibiotic resistance to otherwise susceptible bacteria. As the clinics gradually embrace genome sequencing for microbiological analyses, the ability to use genomic sequencing data to predict relevant phenotypes such as antibiotic resistance will be essential and desirable. This study serves as a foundation for the development of future technologies that could utilize genomic sequencing to analyze the molecular epidemiological trends for bacterial strains reliably, and provide an early-warning measure for the edge towards antimicrobial resistance, crucially informing on clinical intervention strategies.

Result and discussion

Pneumococcal strains and phenotypes of antibiotic resistance

We analyzed 1,680 *S. pneumoniae* isolates systematically selected from diverse sources as follows: 350 invasive isolates from Nijmegen, the Netherlands [19]; a collection of 680 and 332 systematically selected nasopharyngeal carriage (NP) isolates from Massachusetts, USA [16] and Maela, Thailand respectively [14, 17]; and 318 isolates from pediatric suffering from SCD in the USA [18]. Minimum inhibitory concentration breakpoints from the Clinical and Laboratory Standards Institute (CLSI) 2008; and where CLSI breakpoints were unavailable, guidelines from the European Committee on Antimicrobial Susceptibility Testing (EUCAST) were applied to separate the resistant and susceptible isolates accordingly. Compared to the general population, SCD patients are usually at high risk of contracting potentially fatal IPD. Therefore, they receive long-term antibiotic prophylaxis and frequent empiric antibiotic treatment. In response to the antibiotic selective pressure, pneumococci isolated from SCD patients have been shown to exhibit high rates of antibiotic resistance [18, 20]. For carriage isolates from Maela and Massachusetts, 54 of the 1,012 were resistant to trimethoprim, penicillin, erythromycin and cotrimoxazole, representing about 0.054% resistance to four classes of essential antibiotics (multidrug resistance; MDR). Additionally, all 263 carriage isolates that showed full resistance to penicillin were also resistant to at least one other antibiotic of a different class (~26% resistance to penicillin and one other antibiotic). In contrast, only one isolate from Nijmegen (10208_2#41) showed resistance to all antibiotics tested

except for tobramycin (~ 0.0029% MDR). Of the three penicillin-resistant isolates from Nijmegen, one (9953_7#71) also exhibited resistance to tobramycin. A total of 1,012 pneumococcal carriage isolates (Massachusetts and Maela), and 350 IPD isolates (Nijmegen) were included in the variants screening GWAS. The 318 SCD isolates were included in the post-association evaluation of SNPs to determine the distance to resistance because SCD patients are frequently treated with antimicrobials and were expected to exhibit substantial resistance that could skew the associations. Generally, the Nijmegen cohort had the lowest percentage proportion of isolates showing antibiotic resistance; penicillin 0.86% (3 isolates), trimethoprim 4.29% (15), erythromycin 2.28% (8), and cotrimoxazole 4.29% (15), as compared to the selection of Maela and Massachusetts isolates; penicillin 25.99% (263), trimethoprim 14.13% (143 isolates), erythromycin 35.57% (360), and cotrimoxazole 27.76% (281). Surprisingly, the proportions of resistance to fluoroquinolones; ciprofloxacin 22.86% (80) and ofloxacin 50% (175), as well as the aminoglycoside tobramycin 49.14% (172) were remarkably high in the Nijmegen IPD isolates (Figure 1; Supplementary Table 1). Tobramycin could especially be used as a combination drug for selective decontamination of the digestive tract in critically ill patients [21] but not for IPD control in the Netherlands. That large selection for pneumococcal resistance to tobramycin may, therefore, be due to extensive use of the antibiotic or as a result of pleiotropy or linked resistance with other classes of antibiotics.

Correcting for population stratification in bacterial GWAS

S. pneumoniae is highly recombinogenic leading to strains with very diverse genomes [17, 22]. Besides exchanging DNA fragments within pneumococci and other viridian *Streptococci*, the nasopharynx also houses other commensal microorganisms, such as *Haemophilus influenzae*, *Moraxella catarrhalis*, and *Staphylococcus aureus*, which could readily provide genetic material for recombination. Although affecting smaller parts of the genome and not occurring in every generation, the high frequency of homologous recombination is advantageous in creating genetic admixture into bacterial populations in a manner similar to sexual reproduction in humans. This breaks up the strong linkage disequilibrium (LD; large haplotype blocks) which usually muddle bacterial GWAS and help to identify associations unlikely to have occurred by chance [14]. The confounding effect of the clonal population structure is, in this sense, restrained in highly recombinogenic bacteria like *S. pneumoniae*. However, the different cohorts we sampled provided strains with multi-lineage clonal backgrounds whose population structure may lead to false associations [13], particularly since resistant clones may have a fitness advantage in settings where antibiotics use is frequent. Therefore, we corrected for population stratification using the genetic subpopulations (represented by the sequence clusters; SCs) determined using BAPS and/or the method proposed by Prosperi *et al.* [23] (see Materials and methods), allowing for precise discrimination of causal variants from linked variants. This approach and the large dataset we sampled enabled for reliable

identification and separation of actual phenotypic correlates of antibiotic resistance from confounders of the clonal population.

Specific polymorphisms and genes strongly associated with antibiotic resistance

Initial candidate putative causal variants were selected as SNPs showing statistically significant associations, *p-values* < 0.01, controlled for population substructure and Bonferroni-adjusted for multiple testing. Separate associations were tested for resistance to penicillin, trimethoprim, cotrimoxazole, erythromycin, ciprofloxacin, ofloxacin, and

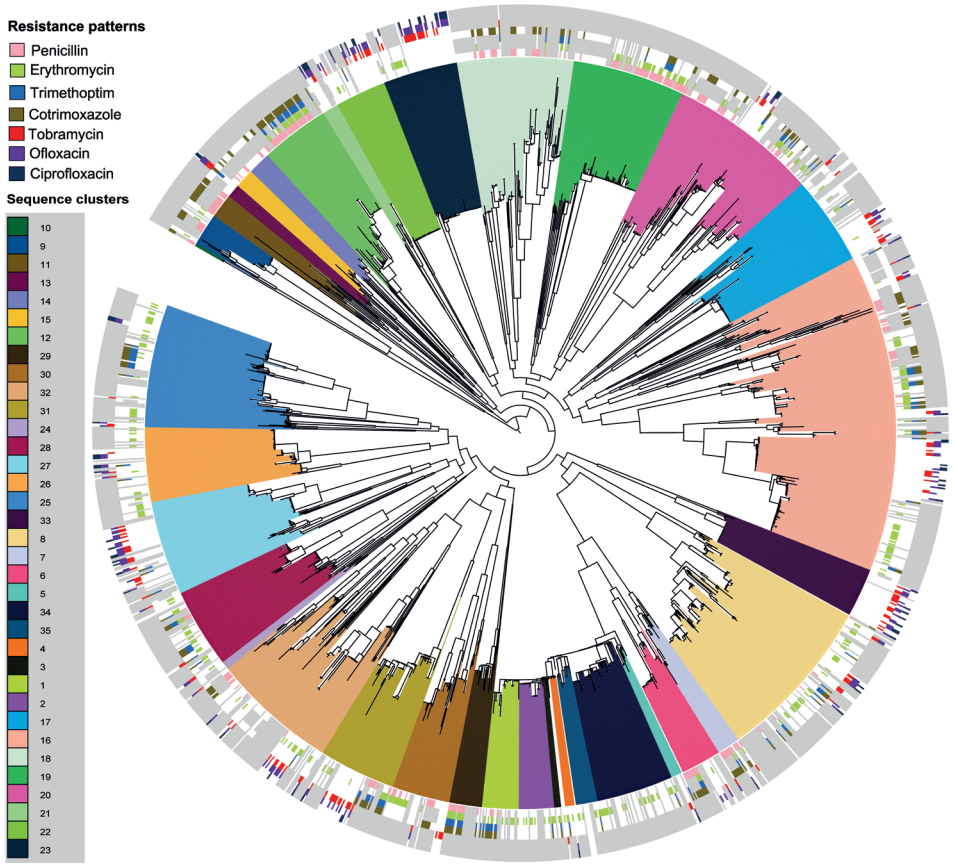


Figure 1. A maximum likelihood phylogeny of the concatenated variant regions from the core genome of 1682 pneumococcal isolates. The clades are colored according to the sequence clusters. The circular stripes represent the antibiotic resistance phenotypes starting from the inside: pink; penicillin, green; erythromycin, blue; trimethoprim, gold; cotrimoxazole, red; tobramycin, purple; ofloxacin, and navy-blue; ciprofloxacin.

Table 1. Single nucleotide polymorphisms that are determining of penicillin resistance

BP Position	RF importance	RF *p*-value	CMH *p*-value	Annotation	Gene
1613086	20.62957785	0	3.308E-99	penicillin-binding protein 2b	*penA*
1612897	19.83961807	0	6.649E-85	penicillin-binding protein 2b	*penA*
333792	11.86365577	8.69E-98	1.234E-99	Holliday junction-specific endonuclease	*recU*
333282	11.70867819	1.55E-18	5.304E-55	penicillin-binding protein 1A	*pbp1A*
334107	11.29654702	7.83E-162	9.151E-72	Holliday junction-specific endonuclease	*recU*
334639	10.42365322	0	1.203E-68	hypothetical protein	-
294991	10.24670847	0	5.276E-60	phospho-N-acetylmuramoyl-pentapeptide-transferase	*mraY*
335104	8.53526957	3.3E-17	2.659E-68	DivIVA protein	-
333345	7.379638353	6.46E-16	8.607E-24	penicillin-binding protein 1A	*pbp1A*
1613422	7.320298527	9.51E-56	4.344E-55	penicillin-binding protein 2b	*penA*
292563	7.167180495	8.14E-08	1.294E-55	penicillin binding protein 2x	*pbpX*
332247	6.771832321	4.27E-26	3.671E-55	penicillin-binding protein 1A	*pbp1A*
335955	6.631966798	8.25E-35	9.742E-45	RNA methylase family protein	-
1613770	6.108715366	7.27E-87	2.023E-78	penicillin-binding protein 2b	*penA*
333386	5.891738108	4.68E-13	8.607E-24	penicillin-binding protein 1A	*pbp1A*
1532326	5.159428663	2.95E-11	4.143E-72	dihydrofolate reductase	*dyr*
1531915	5.154251817	1.21E-11	1.717E-20	ATP-dependent protease ATP-binding subunit ClpX	*clpX*
295737	4.619274035	1.67E-75	7.271E-49	ATP-dependent protease ATP-binding subunit ClpL	*clpL*
296748	4.473732826	4.15E-08	2.701E-36	ATP-dependent protease ATP-binding subunit ClpL	*clpL*
2193975	4.432372852	3.41E-61	8.905E-27	elongation factor Ts	*tsf*

tobramycin. Penicillin and other β-lactams have for long been the primary means of treating pneumococcal infections and are perhaps the most widely used group of antibiotics that work by inhibiting bacterial cell wall biosynthesis [24]. Resistance arises when the organism produces beta-lactamases, which are enzymes that cleave the beta-lactam ring, or modifies the drug targets; 'penicillin binding proteins'; PBPs [25]. Published research implicates changes in PBPs as the primary determinants of β-lactams resistance [26-30]. Variations in PBP2b and PBP2x modulate low-level (intermediate) β-lactams resistance with additional changes in PBP1a leading to extreme resistance. Growth inhibition assays have also shown that β-lactams primarily kill the pneumococcus by inhibiting PBPs, particularly PBP2x [31]. PBP transpeptidase signatures have also been shown to be significant indicators of resistance levels in various β-lactams [32]. SNPs associated with beta-lactam resistance were previously reported using GWAS [14]. In our analysis, however, we identified a significantly higher number of SNPs that confer resistance to penicillin (4,137) than Chewapreecha and colleagues (858 and 1,721 in Maela and Massachusetts cohorts respectively - 301 common SNPs). The difference could be explained by the fact that Chewapreecha *et al* replicated their associations in two independent cohorts and only collated candidate SNPs that were identified to be common between the two groups subsequently minimizing false positives. In contrast, we selected all SNPs showing significant statistical association with the phenotype after correction for multiple testing (*p*-value <0.01).

Moreover, we potentially introduced more genotypic variance by analyzing a mixture of isolates, especially the disease isolates, from different geographical areas. The divergent genotypes in practice resulted into more sequence clusters (used for population substructure stratification) further partitioning the phylogenetic clusters/clades that may have been considered similar in the Chewapreecha study. Therefore, our approach allowed for identification of more SNPs. Of the 4,137 SNPs associated with resistance to penicillin, 3,589 were in coding or intragenic sequences and 548 in non-coding or intergenic sequences (Figure 2; Supplementary Table 2). Interestingly, a q-q plot of the GWAS results shows a sharp deviation above an expected *p*-value indicating the presence of unusually high linkage disequilibrium and strong association of the phenotype with SNPs in heavily genotyped loci.

To further control for inflation and increase confidence in verity, a more stringent *p*-value cut-off was applied. We generated a q-q plot of the penicillin association *p-values* and placed a limit threshold at the point of deviation of the observed *p-values* from the expected *p*-values. This new cut-off (*p*-value<1.5E-20) yielded a smaller subset of 426 SNPs associating to the penicillin resistance phenotype. These SNPs are localized primarily in genes which have been previously reported to be involved in development of penicillin resistance, including genes involved in the peptidoglycan biosynthesis pathway like penicillin binding proteins; PBPs (*pbp1A, pbp1B, pbpX, pbp2A, penA*), peptidoglycan biogenesis transferases (*mraW*, *mraY*) and synthesis of peptidoglycan precursors (*murM*, *murN*), pneumococcal surface protein (*pspA, pspC*), recombination pathway (*recU*), cell division pathway (*gpsB*, *ftsL*), MDR proteins and drug efflux pumps (*pmrA*), drug antiporters [14], heat shock proteins/chaperones (ClpL) [33], and genes implicated in resistance to other essential classes of antibiotics like dihydrofolate reductase (*dyr*); involved in trimethoprim resistance [34, 35].

In the case of penicillin resistance, a single SNP is not enough to confer full resistance. Therefore, we employed a machine learning method, Random Forest (RF) to investigate the combinatorial effect of certain SNPs. The RF model identified a subset of 34 unique SNPs that are predictive of penicillin resistance (Table 1). A GWAS on presence or absence of individual genes also revealed that presence of variants of the cell division initiation protein, gpsB (og_2891 and og_1645; *p*-values 4.73E-44 and 4.06E-43 respectively- Bonferroni-adjusted for multiple testing) significantly correlates with penicillin resistance. Despite being putatively essential and thus expected to be present in all isolates [36], there appears to be two variants of the *gpsB* gene og_1645; n=1457, and og_2891; n=215, judging from their different amino acid sequences. A separate RF analysis on both SNPs and OGs also determined that og_2891 and og_1645 are the only genes among the top 20 features that are predictive of penicillin resistance (Table 2).

Table 2. Single nucleotide polymorphisms and genes that confer penicillin resistance in the pneumococcus

Feature	RF importance	RF *p*-value	CMH *p*-value	Annotation	Gene
1613086	17.19240972	0	3.308E-99	Penicillin-binding protein 2b	*penA*
1612897	16.25074593	0	6.649E-85	Penicillin-binding protein 2b	*penA*
0g_2891	10.99276087	0	4.73E-44	Cell division initiation protein (SP_0372)	*gpSB*
0g_1645	10.67477573	0	4.06E-43	Cell division initiation protein (SP_0372)	*gpSB*
333792	9.977205881	8.69E-98	1.234E-99	Holliday junction-specific endonuclease	*recU*
334107	9.725165692	7.83E-162	9.151E-72	Holliday junction-specific endonuclease	*recU*
333282	9.339362915	1.55E-18	5.304E-55	Penicillin-binding protein 1A	*pbp1A*
334639	8.058415682	0	1.203E-68	Hypothetical protein	-
294991	8.007530946	0	5.276E-60	Phospho-N-acetylmuramoyl-pentapeptide-transferase	*mraY*
335104	7.818192276	3.3E-17	2.659E-68	DivIVA protein	-
292563	6.279875395	8.14E-08	1.294E-55	Penicillin binding protein 2x	*pbpX*
1613422	5.834807369	9.51E-56	4.344E-55	Penicillin-binding protein 2b	*penA*
332247	5.797588876	4.27E-26	3.671E-55	Penicillin-binding protein 1A	*pbp1A*
333345	5.576828511	6.46E-16	8.607E-24	Penicillin-binding protein 1A	*pbp1A*
1613770	5.573167922	7.27E-87	2.023E-78	Penicillin-binding protein 2b	*penA*
3333386	5.202912551	4.68E-13	8.607E-24	Penicillin-binding protein 1A	*pbp1A*
335955	5.175277829	8.25E-35	9.742E-45	RNA methylase family protein	-
1531915	4.829698691	1.21E-11	1.717E-20	ATP-dependent protease ATP-binding subunit ClpX	*clpX*
1532326	4.61272854	2.95E-11	4.143E-72	Dihydrofolate reductase	*dyr*
296748	4.294101416	4.15E-08	2.701E-36	ATP-dependent protease ATP-binding subunit ClpL	*clpL*
296367	4.031726526	7.28E-23	8.841E-38	ATP-dependent protease ATP-binding subunit ClpL	*clpL*

This observation is perhaps a reinforcement that resistance to β-lactams is primarily driven mutations in key genes, and possibly by the presence of certain resistance genes. GpsB is thought to be vital for peripheral and septal peptidoglycan synthesis in *S. pneumoniae*, particularly in the recruitment of PBP1 to the division complex and its removal from the cell pole soon after pole maturation is completed [36]. It shows overlapping, although non-identical, pattern of co-localization with FtsZ during cell division. Depletion of *gpsB* causes division defect characterized by significant cell elongation and enlargement, several unconstricted rings of division proteins Pbp2x, Pbp1a, FtsZ, and MreC, cessation of growth, and eventually cell lysis in *S. pneumoniae* D39 [36]. These phenotypes are similar to those observed in Pbp2x depletion [37] or inhibition of Pbp2x by the β-lactam antibiotic methicillin [36]. Therefore, the observed association is most likely a secondary effect of the functions of gpsB in peptidoglycan synthesis during cell division which may have fitness and pleiotropic consequences in maintaining cell integrity rather than a direct role in resistance.

Penicillin-resistant pneumococcus exhibit varying patterns of resistance to other β-lactams and are generally resistant to other classes of antibiotics that are usually active against pneumococci [38-40]. We evaluated resistance to trimethoprim, cotrimoxazole, erythromycin, tobramycin, ofloxacin and ciprofloxacin. Trimethoprim and cotrimoxazole (a combination of trimethoprim and sulfamethoxazole) reduce the ability of some

bacteria to utilize folic acid for growing [34], by blocking folate metabolism via *dhfR* or *dyr* (encoding dihydrofolate reductase) and *folP/sulA* (encoding dihydropteroate synthase) respectively. These drugs interrupt two crucial steps required in the biosynthesis of bacterial proteins. Mutational or recombinational changes on the target enzymes; *dhfR* and *folP* or their promoter regions have been reported to enhance resistance to trimethoprim and cotrimoxazole [34, 35, 41, 42]. We identified SNPs associated with resistance to trimethoprim and cotrimoxazole in various genes encoding enzymes involved in folate metabolism, including *dyr*, *folE*, and *folP* (Figure 2; Supplementary Tables 3-4), and SNPs in genes implicated in resistance to other essential antibiotics like penicillin. RF analysis revealed that only mutations in folate metabolism (*dyr, folC, folE, folP,* and 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase; HPPK (EC 2.7.6.3) encoded by *folK or sulD*), and chaperones/ATP-dependent proteases (*clpL*, *clpX*) [43], as well as linked mutations in PBPs (*pbp1A, pbpX, penA*), recombination proteins (*recR, recU*), and peptidoglycan biogenesis transferases (*mraW*, *mraY*), were predictive of cotrimoxazole and trimethoprim resistance (Tables 3 and 4). Also, there is a very strong co-selection for resistance to other classes of antibiotics, especially β-lactams. However, it is not clear whether this is because of frequent simultaneous use of these antibiotics, or an epistatic effect on linked compensatory pathways.

Erythromycin and other macrolides inhibit protein synthesis by penetrating the bacterial cell membrane and binding to the ribosomal RNA molecules, particularly in the 50S subunit, of the bacterial ribosome blocking the exit of the growing peptide chain. Macrolides remain an important class of antibiotics for pneumococcal disease. In the USA, macrolides are used as monotherapy for outpatient pneumonia and in combination with β-lactams for more severe pneumonia [44]. The macrolide azithromycin is combined with ceftriaxone as empiric therapy for severe pneumonia, and clarithromycin is a second line treatment for mild community-acquired pneumonia in Australia [45]. However, the prevalent use of these antibiotics in other indications like non-pneumococcal respiratory tract infections [46], trachoma and sexually transmitted diseases [47], and chronic obstructive pulmonary disease (COPD) [48] may be the primary driver of selection for macrolide resistance in pneumococci. Pneumococcal resistance to macrolides is caused by drug efflux or alteration of the target site [49-53]. The phenotypic expression of target-site modification can be inducible or constitutive [49], and can be confirmed with the presence of *mefA/E* and *ermB* genes. *MefA* and *mefE* share >90% sequence homology and are carried in transposons which are comprised of additional open reading frames [54]. Before correcting for multiple testing, we observe a statistically significant association between the presence of *ermB* (og_1123) and resistance to erythromycin (*p*-value 3.137E-05).

Figure 2. Manhattan plots summarizing the statistical significance of genome-wide associations between whole-genome SNPs and resistance to various antibiotics. Specific loci that are significantly associated with resistance to antibiotic are shown in the top panels.

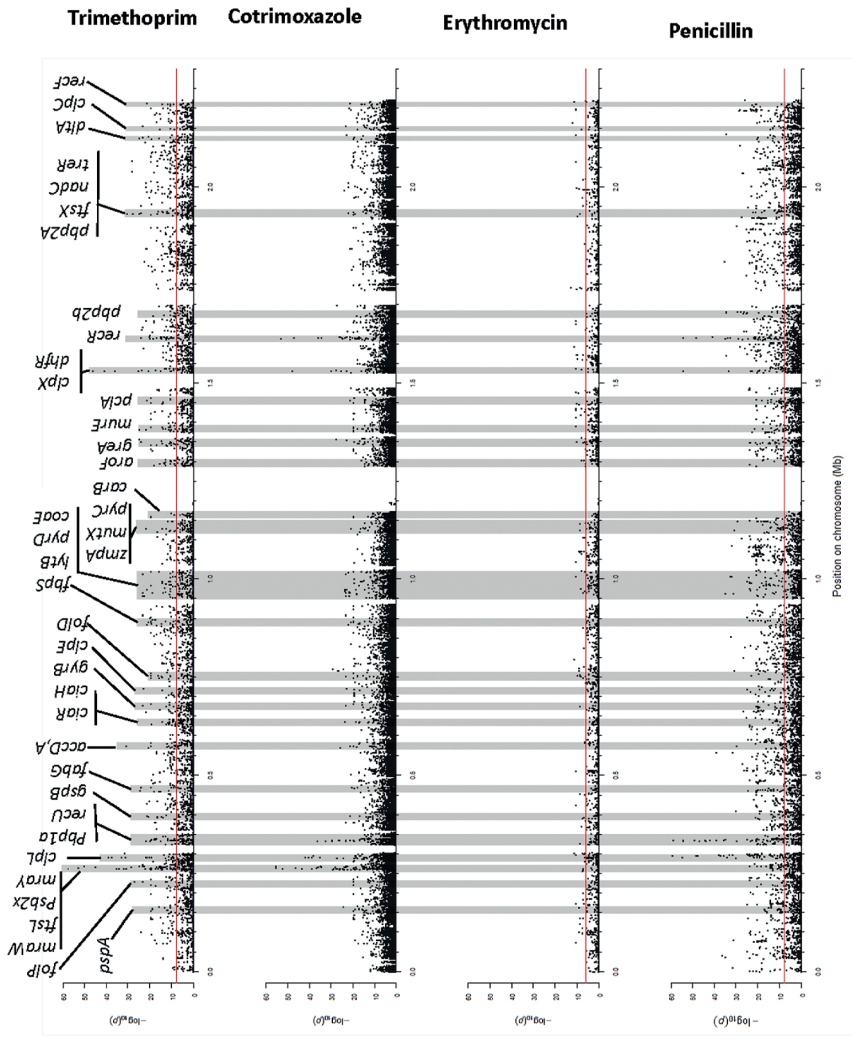


Table 3. Single nucleotide polymorphisms and genes that confer cotrimoxazole resistance in the pneumococcus

Feature	RF importance	RF *p*-value	CMH *p*-value	Annotation	Gene
1531915	12.06447624	0	3.65E-32	ATP-dependent protease ATP-binding subunit ClpX	*clpX*
293661	10.58122979	2.05E-112	1.3E-75	penicillin binding protein 2x	*pbpX*
267233	10.09698603	0	5.656E-82	GTP cyclohydrolase I	*folE*
1532245	9.521815356	0	2.616E-81	dihydrofolate reductase	*dyr*
1613086	9.414234472	5.99E-237	8.051E-62	penicillin-binding protein 2b	*penA*
264912	8.905643954	0	1.032E-75	dihydropteroate synthase	*folP*
1531651	7.923116586	0	4.595E-60	ATP-dependent protease ATP-binding subunit ClpX	*clpX*
265536	7.247622768	5.56E-86	3.638E-67	dihydropteroate synthase	*folP*
291557	6.341493841	3.03E-43	3.097E-61	cell division protein	*ftsL*
292017	5.849575748	7.35E-31	4.103E-50	penicillin binding protein 2x	*pbpX*
1532054	5.822732316	0	1.203E-28	hypothetical protein	-
1612897	5.726042635	8.3E-102	1.213E-49	penicillin-binding protein 2b	*penA*
262352	5.671693703	0	1.208E-38	-	-
263885	5.610620269	9.89E-132	7.608E-41	permease	-
332799	5.455168203	5.73E-56	1.758E-48	penicillin-binding protein 1A	*pbp1A*
262539	5.429029061	0	2.304E-33	-	-
263190	5.333550128	9.15E-144	4.673E-26	permease	-
1613770	5.331345531	3.4E-65	2.497E-51	penicillin-binding protein 2b	*penA*
291982	5.096429153	4.18E-34	3.797E-47	penicillin binding protein 2x	*pbpX*
335104	4.434933551	0.000000165	1.249E-40	DivIVA protein	-

The OG comprises of two variants of the "23S-rRNA (adenine(2058)-N(6)) methyltransferase" (ermB): one variant is of the *Staphylococcus aureus* origin (ungapped blastP alignment with 100% sequence identity over 100% sequence coverage and an e-value of 3e-176). Surprisingly, this association significance wanes after correcting for multiple testing (*p*-value 0.09; Bonferroni-corrected for multiple testing), perhaps due to the low prevalence of *ermB* (n=124). Nonetheless, we identified polymorphisms in the 16S rRNA; *rsmE*, 50S rRNA; *rplM*/*S*/*B*/*T*, *rpmA*/*E2*/*F*/*H*, *rplE*/L/I, and 30S rRNA; *rpsA*/*M*/*N*/*P*/*D*/*H* molecules among others, that significantly associate with erythromycin resistance (Figure 2; Supplementary Table 5). In the RF analysis, the presence of various genes was associated with resistance to erythromycin (Table 5). They include the macrolide efflux pump *mefA* (og_1312): [54]; ImpB/MucB/SamB family protein (og_1652): a family of error-prone DNA polymerases involved in DNA repair [55]; the YolD-like protein (og_1379), a group of functionally uncharacterized proteins predicted to be functionally alike to the UmuD subunit of polymerase V from Gram-negative bacteria [56] and the ribose import ATP-binding protein (og_791; SP_1114). The functional protein interaction network STRING shows that the ribose import ATP-binding protein (og_791; SP_1114) directly interacts with various efflux pumps associated with antibiotic resistance: the drug efflux ABC transporter ATP-binding protein/permease (SP_1342), the MATE efflux pump (SP2065), and the MATE family DinF transport system (SP1939) [57]. UmuD [58] is

Table 4. Single nucleotide polymorphisms and genes that confer trimethoprim resistance in the pneumococcus

Feature	RF *p*-value	CMH *p*-value	Annotation	Gene
267970	4.39E-262	2.51E-13	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	-
291557	4.19E-57	6.03E-15	Cell division protein	*ftsL*
292017	4.99E-52	2.89E-16	Penicillin binding protein 2x	*pbpX*
291982	9.37E-47	1.34E-16	Penicillin binding protein 2x	*pbpX*
cog_1652	3.01E-198	3.54E-08	ImpB/MucB/SamB family protein	-
cog_1312	1.79E-58	0.00000173	Macrolide efflux pump	-
cog_1379	4.39E-85	0.000000119	YolD-like protein	-
cog_791	3.88E-48	0.00000049	Ribose import ATP-binding protein rbsA	-
cog_2061	1.41E-90	0.000000119	Hypothetical protein	-
291286	3.39E-15	7.68E-17	S-adenosyl-methyltransferase MraW	*mraW*
290545	3.38E-26	6.7E-12	-	-
1558235	4.29E-19	4.69E-11	Cytidylate kinase	*cmk*
cog_1645	0	0.001967	Cell division initiation protein	-
cog_2891	0	0.000622	Cell division initiation protein	-
cog_182	1.49E-25	0.001306	Binding-protein-dependent transport systems inner membrane component	-
2002234	2.83E-28	6.6E-15	PTS system ascorbate-specific transporter subunit IIC	-
1459029	5.58E-30	4.56E-12	Asparaginyl-tRNA synthetase	*asnC*
887794	1.4E-27	4.56E-12	DNA polymerase III subunit epsilon	-
1800425	0.0000387	1.73E-21	DegT/DnrJ/EryC1/StrS family amino sugar synthetase	-

involved in UV protection and mutation in Gram-negative strains. It may modify the DNA replication machinery to allow bypass synthesis across a damaged template. ImpB copies undamaged DNA at stalled replication forks, which arise *in vivo* from mismatched or misaligned primer ends. These misaligned primers could be extended by DNA polymerase IV subunit (polIV).

Fluoroquinolones; Ofloxacin and ciprofloxacin inhibit bacterial DNA synthesis by promoting cleavage of bacterial DNA in the DNA-enzyme complexes of DNA gyrase and type IV topoisomerase; more specifically, the *gyrA* and *gyrB* subunits of DNA gyrase, and *parC* and *parE* subunits of topoisomerase IV, resulting in rapid bacterial death. Fluoroquinolone activity in Gram-positive bacteria usually results from inhibition of DNA type IV topoisomerase whereas activity in Gram-negative bacteria corresponds with inhibition of DNA gyrase. Ciprofloxacin binds exclusively with topoisomerase IV whereas ofloxacin binds more avidly with topoisomerase IV and also binds gyrase. Resistance in *S. pneumoniae* to fluoroquinolone is caused predominantly by mutations in *gyrA* or *parC*, and occasionally in *parE*, which reduces binding of the drug to the site of activity [59-63]. Pneumococcal resistance to fluoroquinolone is thought to be a stepwise process; most intermediately resistant strains accumulate first-step mutations, which usually involve only a single mutation in the target genes [59], although, these strains do tend to go on to develop subsequent second-step mutations which significantly diminishes the activity of most fluoroquinolones and renders the strains highly resistant [64].

Table 5. Single nucleotide polymorphisms and genes that confer erythromycin resistance in the pneumococcus

Feature	RF importance	RF *p*-value	CMH *p*-value	Annotation	Gene
og_1652	7.180656298	0	3.701E-14	ImpB/MucB/SamB family protein	-
og_1379	7.175992284	0	6.409E-15	YolD-like protein	-
og_2061	7.097377018	0	6.409E-15	hypothetical protein	-
og_1312	6.485953794	1.71E-217	2.683E-14	Macrolide efflux pump	*mefA*
og_791	6.017699918	1.03E-210	6.671E-14	Ribose import ATP-binding protein rbsA	-
1741683	4.904707263	8.7E-52	3.01E-23	Nudix-related transcriptional regulator NrtR	-
og_243	4.114250972	0.00000718	0.0002367	IS1380-Spn1 transposase	-
og_10	3.24238261	0.000118	0.01343	3-ketoacyl-ACP reductase	-
290545	3.240769624	1.2E-76	6.7E-12	Tn916 ORF16 ATP/GTP-binding protein	-
297653	3.016787219	4.01E-28	2.24E-16	methyltransferase small domain superfamily	-
og_53	2.971456253	5.89E-142	4.8E-09	UDP-glucose 6-dehydrogenase	-
og_129	2.962230108	0.000564	4.443E-07	chlorohydrolase	-
678325	2.695815692	1.6E-53	2.15E-15	transposase, ISSmi4	-
136082	2.412872116	2.23E-15	7.73E-14	cytidine deaminase	*cdd*
637379	2.41015196	6.07E-44	1.36E-10	nucleotidyl transferase WchZ	-
987093	2.253968194	7.1E-12	4.43E-23	Abi-alpha protein	-
2058103	1.964208468	8.09E-10	5.51E-18	transposase	-
og_4190	1.95393053	0.385	0.4235	IS630-Spn1, transposase Orf1	-
1593956	1.93060019	8.68E-09	1.62E-10	Ribose import ATP-binding protein rbsA	-
794795	1.754564135	0.000000235	1.62E-10	hypothetical protein	-

This increasing mutational heterogeneity makes it harder to use GWAS in precisely identifying the role of each mutation in disseminating antibiotic resistance. Surprisingly, we did not find any associations between SNPs in fluoroquinolone target proteins (DNA topoisomerase and DNA gyrase) and fluoroquinolone resistance (Supplementary Table 5). However, we observed a significant statistical association between mutations in a multi-drug resistance efflux pump (*pmrA*) and resistance to ofloxacin (*p*-value 5.06E-05). PmrA is homologous to other well-studied efflux pumps like NorA and Bmr, whose expression leads to reduced susceptibility against several diverse compounds [65], and has previously been shown to be associated with fluoroquinolone resistance in the pneumococcus [66]. More research will be required to ascertain the role of this efflux pump in propagating ofloxacin resistance in the pneumococcus. We also observed that mutations in the heme exporter protein A (*ccmA*) and the recombination factor protein (*rarA*) associated with resistance to the fluoroquinolone ciprofloxacin.

For pathogenic bacteria, heme is the main source of nutritional iron [67]. Being an essential cofactor for many enzymes, iron is found in all kingdoms of life. The pneumococcus mainly uses iron from hemoglobin and heme to support growth [68]. Paradoxically, heme at high concentrations is toxic because of its high redox potential, thus making it a liability to bacteria [69]. Therefore, bacteria control heme uptake by employing various sequestration, degradation, and efflux mechanisms. While the direct role of CcmA in quinolone-resistance is unclear, archetypes of heme-exporter systems that also modulate other phenotypes have been identified in Gram-positive bacteria [70-

72]. In Group A streptococci the heme exporter confers multi-drug resistance [73]. Apart from the ABC-transporter protein PmrA [74], efflux mechanisms of fluoroquinolone resistance are poorly characterized in the pneumococcus. However, efflux could reduce intracellular fluoroquinolone concentrations to sublethal levels furthering the development of resistance-conferring mutations [75]. More experimentation will, therefore, be required to deduce the exact mechanisms behind this novel association. Recombination is the primary source of genome plasticity, generating phenotypic and genetic diversification in bacteria. Unlike in β-lactam resistance where recombination plays an important role [76], fluoroquinolone resistance arises from very specific resistance-determining mutations within the target proteins [77]. Horizontal transfer of fluoroquinolone resistance loci between viridans group streptococci and the pneumococcus has been shown to occur *in vitro* but not *in vivo* [78], with significantly higher rates during asymptomatic carriage than during invasive isolates [79, 80]. Studies have reported a link between fluoroquinolone resistance and evolution of resistance to penicillin and macrolides [81]; both of which could be fostered by recombination. These studies suggest that the observed association between ciprofloxacin resistance and mutations in *rarA* could be an artifact of the linked resistance with other antibiotics. However, since we cannot rule out a novel mechanism involving these mutations, more studies are required to ascertain this observation.

Predicting antibiotic resistance profiles using genome sequencing data

We presupposed that accumulation of particular SNPs leads to a creep towards antibiotic resistance and subsequently full-resistance, we aimed to categorize the strains according to their phenotype of antibiotic resistance using the putative resistance-conferring SNPs identified by GWAS. This analysis helped determine how far the antibiotic susceptible strains are from attaining full-resistance, and what the creep pattern towards resistance is. We identified the partition between isolates by summing up the logarithmic derivatives of odds ratios (*s*OR) for antibiotic resistance-conferring SNPs. This measure was dubbed the "distance to resistance". Antibiotic resistance profiles differed significantly between sampled populations, and between invasive and nasopharyngeal carriage isolates. Rates of antibiotic resistance were generally low in the Netherlands isolates compared to those from Thailand and USA (Figure 3), probably due to the more stringent antibiotic control policies observed in the Netherlands. The WHO provides defined daily dose (DDD) guidelines for antibiotics use in adults [82]. Research has shown that aggregate hospital antibiotic use by DDD in the United States is discordant with the WHO standards [83]. Most studies established that majority of antibiotics were administered in primary care settings to treat infections for which antibiotic therapy is hardly indicated [84-88]. Such trends are likely to have accelerated the pneumococcal selection for resistance in the US population, and could explain the expected higher levels of resistance as compared to, for example, the Dutch population where antibiotic use is vastly prudent.

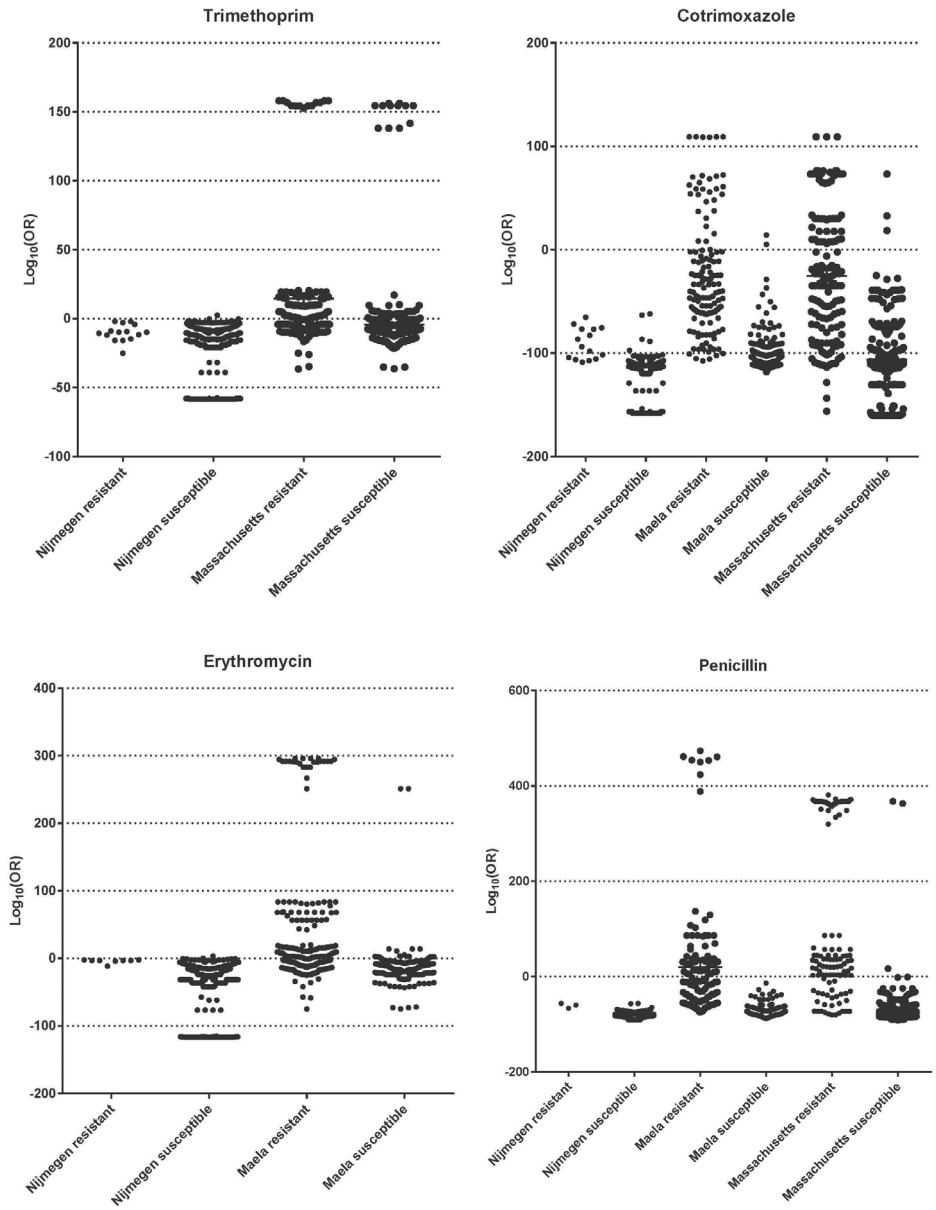


Figure 3. Penicillin, erythromycin, trimethoprim, and cotrimoxazole resistance profiles for isolates from individual geographical locations. Each point represents the cumulative odds ratio effect of SNPs that significantly associate with resistance on each isolate.

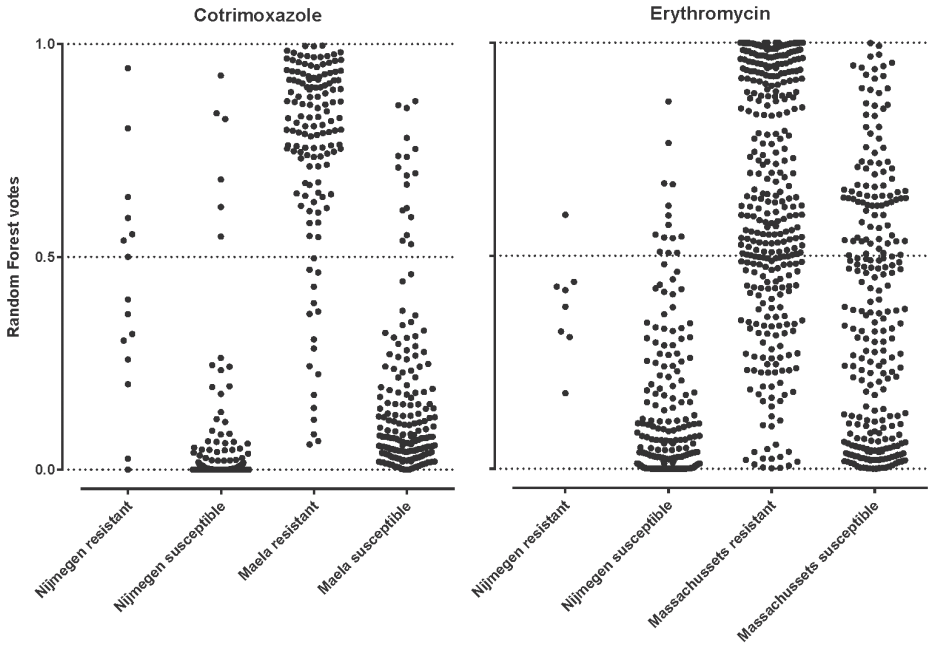


Figure 4. The difference in resistance profiles between invasive and carriage isolates.

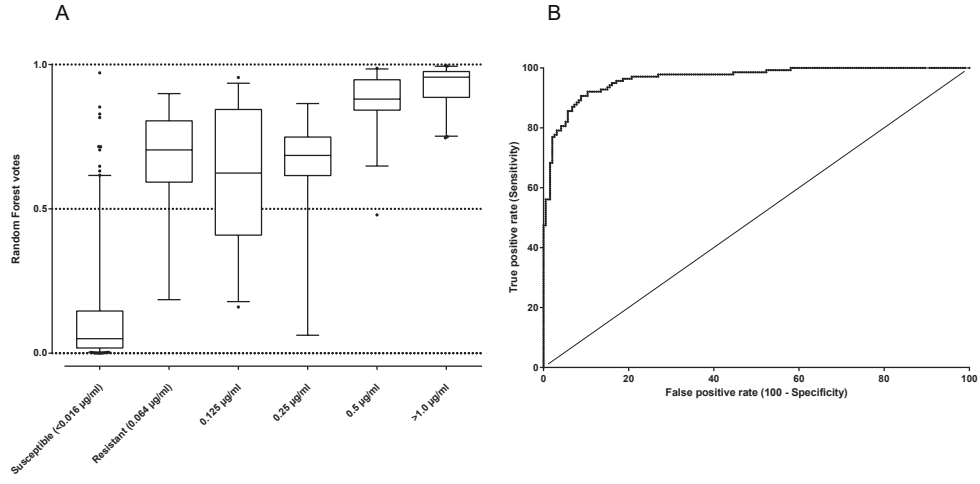


Figure 5. A. A box whisker plot of the random forest votes increasing for resistance relative to the increase in MIC per isolate for penicillin. B. A Receiver Operating Characteristic (ROC) curve showing the discrimination power of the random forest model in predicting penicillin resistance (Area under curve; AUC min= 0.9417 and max= 0.9882 at 99.9% CI).

From this data, it is possible to evaluate the resistance profile of new clinical isolates and to predict strains that are approaching extreme antibiotic resistance using sequence-based approach. We observed that carriage (non-invasive) strains exhibited more resistance to antibiotics than invasive strains: they mostly contained resistance-associated SNPs with high odds ratios (Figure 4). Indeed, there is ample opportunity for pneumococci in carriage to benefit from multiple co-colonizing strains of same or closely related species that provide new genetic material for homologous gene transformation. This phenomenon be could be explained by the fact that rates of pneumococcal transformation are also much higher during colonization than during planktonic growth in sepsis [89].

Pneumococci exhibiting decreased susceptibility to regular penicillin doses are on the rise [38, 39]. Many European guidelines recommend high-dose empirical treatments for systemic infections caused by such strains showing atypical susceptibility. Analyzing the strains for an effective dose, however, relies on laboratory assays. A recent publication reports the use of PBPs transpeptidase signatures (TPDs) from 2,528 clinical pneumococcal isolates to predict the MICs for various β-lactam antibiotics [32]. Li and colleagues constructed predictive models that link amino acid sequence variations in the TPDs of PBP1a, PBP2b, and PBP2x to β-lactam MIC levels among invasive pneumococcal isolates. They identified 68, 78, and 118 unique TPD amino acid sequences for PBP1a, PBP2b, and PBP2x, respectively. Using 307 unique combinations of these sequences which defined the PBP types, they observed that isolates whose PBP types exhibited more than 10% amino acid sequence divergence from a usual susceptible PBP type were associated with increased β-lactam MICs. Such studies raise the optimism that genomic sequencing-based approaches could soon be used as alternatives to phenotypic susceptibility testing. We observed that increase in minimum inhibitory concentrations (MIC) is characterized by an increase in *s*ORs (Figure 5). Additionally, the antibiotics-susceptible isolates (MIC <0.016 μg/ml) have also acquired some level of mutation that brings them closer to low-MIC penicillin resistance, which in clinical practice can be managed by increased penicillin dose. Overall, we could reliably test for decreasing sensitivity to high doses of penicillin treatment based only on genome sequences with this approach.

Prospective value of the distance to resistance

In response to environmental stress over time, bacteria evolve genetic adaptations such as the acquisition of resistance genes or accumulation of critical mutation that confer antibiotic resistance. Figure 6 shows the resistance profiles for four antibiotics over time in different cohorts. There seems to be an increase, however subtle, in the number of isolates that have accumulated more resistance-conferring SNPs in each cohort every year. The IPD isolates from Nijmegen have particularly accumulated combinations of SNPs that confer resistance but are more close to susceptibility as compared to carriage

isolates. The onset of these SNPs is perhaps indicative of the development of antibiotic resistance. However, more studies will be required to confirm this observations. Put together, these results show that by starting with a curated reference database of genes and SNPs/alleles that confer resistance in historical and contemporary isolates; it is possible to model a framework that predicts a creep towards resistance over time and extreme antimicrobial resistance for various antibiotics in new bacterial isolate. Prospecting these changes by looking solely at genome sequence data provides an early warning sign that could significantly benefit public health surveillance.

In conclusion, this study establishes a proof of concept measure for the 'distance to antimicrobial resistance.' Although causality cannot always be drawn from associations, bacterial GWAS provide the means of identifying genomic variants, of a rational basis for functional validation, for important microbial phenotypes like antibiotic resistance. By quantifying the aggregate effects of individual resistance-conferring SNPs on the phenotype, we have demonstrated that prediction of advancing antimicrobial resistance could be achieved *in silico* using genomic sequencing data. Therefore, this study invokes a change of perspective for future research to focus on detecting genetic variants and variations in genetic loci responsible for heralding the creep towards antibiotic resistance in pathogenic bacteria. Such sequencing-based frameworks are not only affordable and consistent but also allow for simultaneous discovery of essential pneumococcal features such as serotype and sequence type. Altogether, this knowledge will greatly inform the choice of clinical intervention and improve public health surveillance thus precluding outbreaks caused by emerging multidrug resistant strains.

Materials and methods

Strains and phenotypes in study

This study included 1,680 pneumococcal isolates and corresponding antibiotic resistance phenotypes; 349 from adults admitted with invasive pneumococcal disease between 2001 and 2011 in two hospitals in Nijmegen, The Netherlands, [19], a systematic selection (three from each "secondary BAPS" cluster) of published carriage isolates from Massachusetts, USA [16] and Maela, Thailand [14, 17], and 318 isolates from children suffering from sickle-cell disease (SCD) in the USA [18] which included isolates from the CDC ABC bacterial surveillance core and published collections [90, 91].

Phenotypes of antimicrobial susceptibility were determined *in vitro* as previously described [15], following the guidelines for antimicrobial susceptibility and minimum inhibitory concentration breakpoints stipulated by the Clinical and Laboratory Standards Institute (CLSI) 2008. Where CLSI guidelines were lacking, the European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines were used. Intermediately

resistant (IR) isolates were grouped as resistant (non-susceptible) when testing for associations.

Determining SNPs and orthologous sequences

Bases were called from mapped sequences using kSNP v2 software [92] against a single reference genome: multidrug-resistant *S. pneumoniae* ATCC 700669; Spain 23F ST81 [93]. A total of 124,310 SNP calls were generated. Filtering for SNPs present in more than ~ 90% of the isolates, (1,500 strains) resulted in 76,429 SNP calls that we used for further analysis. To determine clusters of orthologous sequences, all coding sequences (CDS) from the 1,680 isolates were predicted using Prodigal [94]. All coding sequences were analyzed using USEARCH [95] and aligned in the 'large-scale BLAST score ratio' (LS-BSR) pipeline [96] allowing 10% amino acid difference within clusters. The resulting representative sequences per group ("centroids") were clustered through a Markov Clustering Algorithm (TRIBE-MCL) [97] with an inflation factor of 2.5, resulting into 4687 orthologous groups (OGs). For each OG, we generated a binary metric of the presence (1) or absence (0) of a representative coding sequence(s) (CDS) from each strain. Each strain's contribution of CDS to an OG was subsequently denoted by a single numeric value (1 or 0) to designate the presence or absence of a distinct gene, gene variants, or a group of paralogs. These groups were collated into a binary matrix and formatted for PLINK association analyses [98].

Determining the population structure and performing statistical association

Resistance phenotypes were grouped according to antibiotic classes and analyzed separately for each population, and together in the final association analysis. The population clusters used to control for the effect of clonal inheritance of genetic variants and population stratification were determined using the Bayesian Analysis of Population Structure (BAPS) software [99], and a phylogeny-based partitioning approach as proposed by Prosperi et al. [23], which employs *'ape*, *geiger*, *igraph*, and *phytools'* packages in R software [100]. An alignment of concatenated SNPs from the core or non-repetitive DNA of each of the isolates were analyzed in BAPS as previously described [15]. For the Prosperi clustering, a maximum-likelihood phylogenetic tree was constructed using RAxML version 8.2.0 [101] and an alignment of all the concatenated SNPs from the core genome of all isolates as described before [15]. The general time-reversible model was used to calculate the maximum-likelihood ratios with a γ adjustment for site variation as the nucleotide substitution model. The support for nodes on the tree was tested using a hundred unsystematic bootstrap replicates. Resulting phylogenetic tree was visualized using iTOL version 2.1 [102].

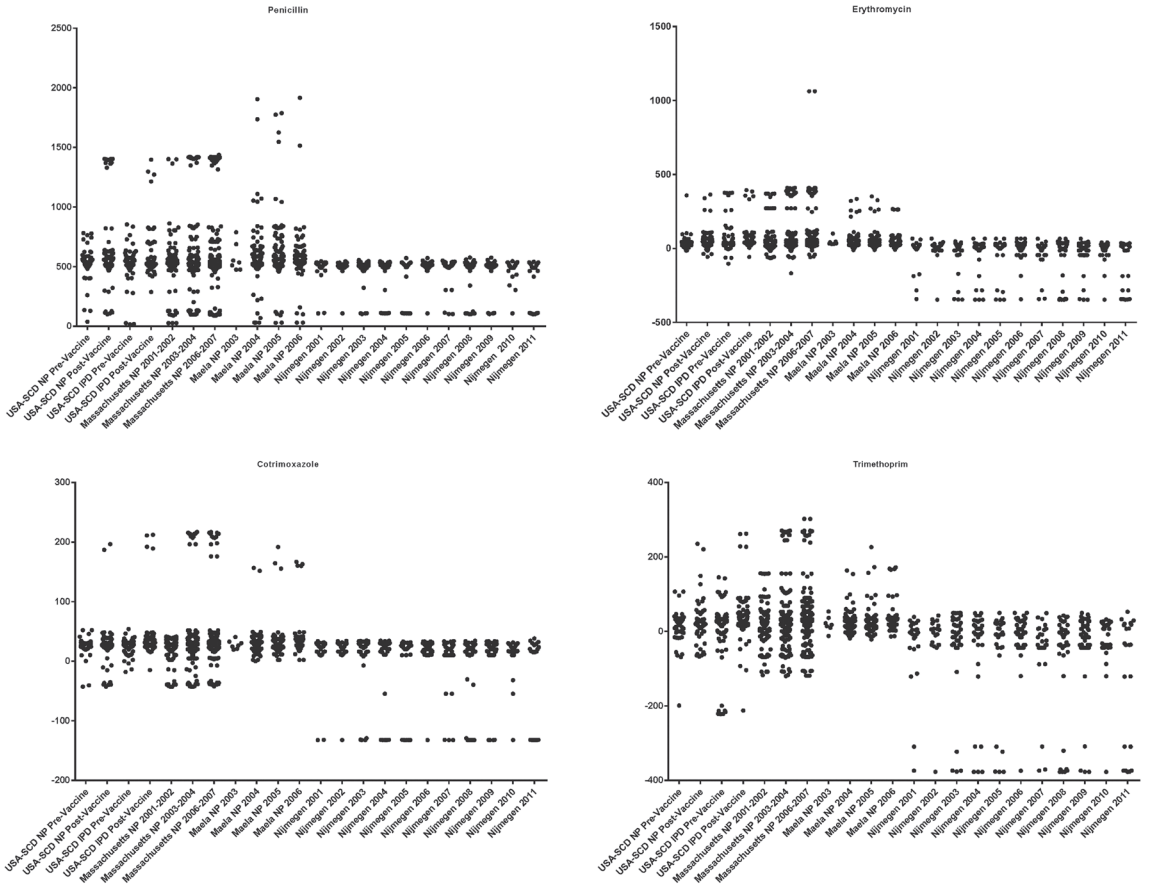


Figure 6.
Variations in the accumulation of resistance-conferring mutations for different antibiotics over time within different cohorts.

We used the Cochran-Mantel-Haenszel (CMH) correlation statistic to test for associations between antibiotic resistance phenotype and SNPs conditional on the population structure. Stratification for population structure minimized falsely positive associations that could be obtained merely by chance. We tested associations for resistance to penicillin, trimethoprim, cotrimoxazole, erythromycin, ofloxacin, ciprofloxacin, and tobramycin. The statistical associations were performed using PLINK software v1.9 [98], and the results visualized as Manhattan and Q-Q plots in R using '*qqman*, *hmisc*, and *ggplot2*' packages. To investigate the effect of SNP combinations, a Random Forest (RF) classification using the Bioconductor *randomForest* package 4.6-10 was performed to discriminate resistant (R) and susceptible (S) isolates. This classification model, consisting of 5000 decision trees was trained on candidate genes and/or SNPs that were determined to be predictive of resistance through GWAS analysis. Statistical significance of the genes or SNPs that were able to discriminate between the classes were calculated by permuting the sample class labels. A normal distribution was determined for mean decrease accuracy (mda) values from the 300 permutations using the *pnorm* package which is part of the R version 3.3.0 distribution. Using the same package a *p*-value was calculated comparing the average of mda values for RF 100 analyses with the original sample classes (to account for slight differences between RF analyses) to the distribution of permuted mda values.

Candidate resistance loci and a measure of the distance to resistance

We selected SNPs showing statistically significant associations (*p-values* < 0.01 at a minor allele frequency > 0.01; Bonferroni-adjusted for multiple testing) as candidates for subsequent analysis. The percentage distribution of these candidate SNPs within resistance isolates relative to the susceptible isolates in each population was computed to determine how they vary in each cohort. For each SNP significantly associated with antibiotic resistance, we determined the odds ratio (OR) and nature (positive or negative) of the correlation. The accumulation of these significant SNPs in each isolate was also defined across all test cohorts. Each SNP present in an isolate was represented by the logarithmic derivative of the odds ratio; $\log_{10}$ (OR): The negative logarithmic values were used for SNPs negatively correlated with resistance. Q-Q plots were used to determine a more stringent *p*-value cut-off. The aggregate effect of the SNPs conferring antibiotic resistance is the sum of all the $\log_{10}$ (OR) values for SNPs above the *p*-value threshold. These represented a measure of the level of resistance of an isolate. These aggregate values were plotted in GraphPad Prism v6.05 software.

Supporting data

Supplementary material for this chapter are available online at:

https://www.dropbox.com/sh/bnx7vad8qivax1d/AABXGROkQDjovRQqmRruQqPOa?dl=0

References

1. Lynch, J.P., 3rd & Zhanel, G.G. *Streptococcus pneumoniae*: epidemiology and risk factors, evolution of antimicrobial resistance, and impact of vaccines. *Curr Opin Pulm Med* **16,** 217-25 (2010).
2. O'Brien, K.L., Wolfson, L.J., Watt, J.P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O.S. & Cherian, T. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *The Lancet* **374,** 893-902 (2009).
3. Zimmerman, R.K. Pneumococcal conjugate vaccine for young children. *Am Fam Physician* **63,** 1991-8 (2001).
4. Singleton, R.J., Hennessy, T.W., Bulkow, L.R., Hammitt, L.L., Zulz, T., Hurlburt, D.A., Butler, J.C., Rudolph, K. & Parkinson, A. Invasive pneumococcal disease caused by nonvaccine serotypes among alaska native children with high levels of 7-valent pneumococcal conjugate vaccine coverage. *JAMA* **297,** 1784-92 (2007).
5. Pelton, S.I., Huot, H., Finkelstein, J.A., Bishop, C.J., Hsu, K.K., Kellenberg, J., Huang, S.S., Goldstein, R. & Hanage, W.P. Emergence of 19A as virulent and multidrug resistant Pneumococcus in Massachusetts following universal immunization of infants with pneumococcal conjugate vaccine. *Pediatr Infect Dis J* **26,** 468-72 (2007).
6. Hansman, D. & Bullen, M.M. A Resistant Pneumococcus. *The Lancet* **290,** 264-265 (1967).
7. Appelbaum, P.C., Bhamjee, A., Scragg, J.N., Hallett, A.F., Bowen, A.J. & Cooper, R.C. *Streptococcus pneumoniae* resistant to penicillin and chloramphenicol. *The Lancet* **2,** 995-7 (1977).
8. Donkor, E.S. & Badoe, E.V. Insights into Pneumococcal Pathogenesis and Antibiotic Resistance. *Advances in Microbiology* **04,** 627-643 (2014).
9. Feil, E.J., Enright, M.C. & Spratt, B.G. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidisand *Streptococcus pneumoniae*. *Research in Microbiology* **151,** 465-469 (2000).
10. Hakenbeck, R., Bruckner, R., Denapaite, D. & Maurer, P. Molecular mechanisms of beta-lactam resistance in *Streptococcus pneumoniae*. *Future Microbiol* **7,** 395-410 (2012).
11. Hsieh, Y.C., Su, L.H., Hsu, M.H. & Chiu, C.H. Alterations of penicillin-binding proteins in pneumococci with stepwise increase in beta-lactam resistance. *Pathog Dis* **67,** 84-8 (2013).
12. Sauerbier, J., Maurer, P., Rieger, M. & Hakenbeck, R. *Streptococcus pneumoniae* R6 interspecies transformation: genetic analysis of penicillin resistance determinants and genome-wide recombination events. *Mol Microbiol* **86,** 692-706 (2012).
13. Chen, P.E. & Shapiro, B.J. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol* **25,** 17-24 (2015).
14. Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E., Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., Turner, P., Bentley, S.D. & Parkhill, J. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet* **10,** e1004547 (2014).
15. Cremers, A.J., Mobegi, F.M., de Jonge, M.I., van Hijum, S.A., Meis, J.F., Hermans, P.W., Ferwerda, G., Bentley, S.D. & Zomer, A.L. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. *Sci Rep* **5,** 14952 (2015).

16. Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P. & Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45,** 656-63 (2013).
17. Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P. & Bentley, S.D. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46,** 305-9 (2014).
18. Carter, R., Wolf, J., van Opijnen, T., Muller, M., Obert, C., Burnham, C., Mann, B., Li, Y., Hayden, R.T., Pestina, T., Persons, D., Camilli, A., Flynn, P.M., Tuomanen, E.I. & Rosch, J.W. Genomic analyses of pneumococci from children with sickle cell disease expose host-specific bacterial adaptations and deficits in current interventions. *Cell Host Microbe* **15,** 587-99 (2014).
19. Cremers, A.J., Meis, J.F., Walraven, G., Jongh, C.E., Ferwerda, G. & Hermans, P.W. Effects of 7-valent pneumococcal conjugate 1 vaccine on the severity of adult 2 bacteremic pneumococcal pneumonia. *Vaccine* **32,** 3989-94 (2014).
20. Adamkiewicz, T.V., Silk, B.J., Howgate, J., Baughman, W., Strayhorn, G., Sullivan, K. & Farley, M.M. Effectiveness of the 7-valent pneumococcal conjugate vaccine in children with sickle cell disease in the first decade of life. *Pediatrics* **121,** 562-9 (2008).
21. Silvestri, L. & van Saene, H.K. Selective decontamination of the digestive tract: an update of the evidence. *HSR Proc Intensive Care Cardiovasc Anesth* **4,** 21-9 (2012).
22. Dagerhamn, J., Blomberg, C., Browall, S., Sjostrom, K., Morfeldt, E. & Henriques-Normark, B. Determination of accessory gene patterns predicts the same relatedness among strains of *Streptococcus pneumoniae* as sequencing of housekeeping genes does and represents a novel approach in molecular epidemiology. *J Clin Microbiol* **46,** 863-8 (2008).
23. Prosperi, M.C., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M.C., Gismondo, M.R., Sighinolfi, L., Gray, R.R., Salemi, M., Zazzi, M., De Luca, A. & group, A.c. A novel methodology for large-scale phylogeny partition. *Nat Commun* **2,** 321 (2011).
24. Elander, R.P. Industrial production of beta-lactam antibiotics. *Appl Microbiol Biotechnol* **61,** 385-92 (2003).
25. Holten, K.B. & Onusko, E.M. Appropriate prescribing of oral beta-lactam antibiotics. *Am Fam Physician* **62,** 611-20 (2000).
26. Dowson, C.G., Johnson, A.P., Cercenado, E. & George, R.C. Genetics of oxacillin resistance in clinical isolates of *Streptococcus pneumoniae* that are oxacillin resistant and penicillin susceptible. *Antimicrob Agents and Chemother* **38,** 49-53 (1994).
27. Grebe, T. & Hakenbeck, R. Penicillin-binding proteins 2b and 2x of *Streptococcus pneumoniae* are primary resistance determinants for different classes of beta-lactam antibiotics. *Antimicrob Agents Chemother* **40,** 829-34 (1996).
28. Hakenbeck, R., Kaminski, K., Konig, A., van der Linden, M., Paik, J., Reichmann, P. & Zahner, D. Penicillin-binding proteins in beta-lactam-resistant *Streptococcus pneumoniae*. *Microb Drug Resist* **5,** 91-9 (1999).
29. Munoz, R., Dowson, C.G., Daniels, M., Coffey, T.J., Martin, C., Hakenbeck, R. & Spratt, B.G. Genetics of resistance to third-generation cephalosporins in clinical isolates of *Streptococcus pneumoniae*. *Mol Microbiol* **6,** 2461-5 (1992).

30. Smith, A.M. & Klugman, K.P. Alterations in PBP 1A essential-for high-level penicillin resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **42,** 1329-33 (1998).
31. Kocaoglu, O., Tsui, H.C., Winkler, M.E. & Carlson, E.E. Profiling of beta-lactam selectivity for penicillin-binding proteins in *Streptococcus pneumoniae* D39. *Antimicrob Agents Chemother* **59,** 3548-55 (2015).
32. Li, Y., Metcalf, B.J., Chochua, S., Li, Z., Gertz, R.E., Jr., Walker, H., Hawkins, P.A., Tran, T., Whitney, C.G., McGee, L. & Beall, B.W. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting beta-Lactam Resistance Levels in *Streptococcus pneumoniae*. *MBio* **7**(2016).
33. Tran, T.D., Kwon, H.Y., Kim, E.H., Kim, K.W., Briles, D.E., Pyo, S. & Rhee, D.K. Decrease in penicillin susceptibility due to heat shock protein ClpL in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **55,** 2714-28 (2011).
34. Huovinen, P. Resistance to trimethoprim-sulfamethoxazole. *Clin Infect Dis* **32,** 1608-14 (2001).
35. Pikis, A., Donkersloot, J.A., Rodriguez, W.J. & Keith, J.M. A conservative amino acid mutation in the chromosome-encoded dihydrofolate reductase confers trimethoprim resistance in *Streptococcus pneumoniae*. *J Infect Dis* **178,** 700-6 (1998).
36. Land, A.D., Tsui, H.C., Kocaoglu, O., Vella, S.A., Shaw, S.L., Keen, S.K., Sham, L.T., Carlson, E.E. & Winkler, M.E. Requirement of essential Pbp2x and GpsB for septal ring closure in *Streptococcus pneumoniae* D39. *Mol Microbiol* **90,** 939-55 (2013).
37. Berg, K.H., Stamsas, G.A., Straume, D. & Havarstein, L.S. Effects of low PBP2b levels on cell morphology and peptidoglycan composition in *Streptococcus pneumoniae* R6. *J Bacteriol* **195,** 4342-54 (2013).
38. Perez-Trallero, E., Garcia-de-la-Fuente, C., Garcia-Rey, C., Baquero, F., Aguilar, L., Dal-Re, R., Garcia-de-Lomas, J. & Spanish Surveillance Group for Respiratory, P. Geographical and ecological analysis of resistance, coresistance, and coupled resistance to antimicrobials in respiratory pathogenic bacteria in Spain. *Antimicrob Agents Chemother* **49,** 1965-72 (2005).
39. Perez-Trallero, E., Marimon, J.M., Gonzalez, A., Vicente, D. & Garcia-Arenzana, J.M. Spectrum of antibiotic resistance of the Spain14-5 *Streptococcus pneumoniae* clone over a 22 year period. *J Antimicrob Chemother* **53,** 620-5 (2004).
40. Abgueguen, P., Azoulay-Dupuis, E., Noel, V., Moine, P., Rieux, V., Fantin, B. & Bedos, J.P. Amoxicillin is effective against penicillin-resistant *Streptococcus pneumoniae* strains in a mouse pneumonia model simulating human pharmacokinetics. *Antimicrob Agents Chemother* **51,** 208-14 (2007).
41. Buwembo, W., Aery, S., Rwenyonyi, C.M., Swedberg, G. & Kironde, F. Point Mutations in the folP Gene Partly Explain Sulfonamide Resistance of *Streptococcus mutans*. *Int J Microbiol* **2013,** 367021 (2013).
42. Haasum, Y., Strom, K., Wehelie, R., Luna, V., Roberts, M.C., Maskell, J.P., Hall, L.M. & Swedberg, G. Amino acid repetitions in the dihydropteroate synthase of *Streptococcus pneumoniae* lead to sulfonamide resistance with limited effects on substrate K(m). *Antimicrob Agents Chemother* **45,** 805-9 (2001).
43. Piotrowski, A., Burghout, P. & Morrison, D.A. spr1630 is responsible for the lethality of clpX mutations in *Streptococcus pneumoniae*. *J Bacteriol* **191,** 4888-95 (2009).
44. Mandell, L.A., Wunderink, R.G., Anzueto, A., Bartlett, J.G., Campbell, G.D., Dean, N.C., Dowell, S.F., File, T.M., Jr., Musher, D.M., Niederman, M.S., Torres, A., Whitney, C.G., Infectious Diseases Society of, A. & American Thoracic, S. Infectious Diseases Society of

America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin Infect Dis* **44 Suppl 2,** S27-72 (2007).
45. Cheng, A.C. & Jenney, A.W.J. Macrolide resistance in pneumococci—is it relevant? *Pneumonia* **8,** 1-3 (2016).
46. Suda, K.J., Hicks, L.A., Roberts, R.M., Hunkler, R.J. & Taylor, T.H. Trends and seasonal variation in outpatient antibiotic prescription rates in the United States, 2006 to 2010. *Antimicrob Agents Chemother* **58,** 2763-6 (2014).
47. Keenan, J.D., Klugman, K.P., McGee, L., Vidal, J.E., Chochua, S., Hawkins, P., Cevallos, V., Gebre, T., Tadesse, Z., Emerson, P.M., Jorgensen, J.H., Gaynor, B.D. & Lietman, T.M. Evidence for clonal expansion after antibiotic selection pressure: pneumococcal multilocus sequence types before and after mass azithromycin treatments. *J Infect Dis* **211,** 988-94 (2015).
48. Hare, K.M., Singleton, R.J., Grimwood, K., Valery, P.C., Cheng, A.C., Morris, P.S., Leach, A.J., Smith-Vaughan, H.C., Chatfield, M., Redding, G., Reasonover, A.L., McCallum, G.B., Chikoyak, L., McDonald, M.I., Brown, N., Torzillo, P.J. & Chang, A.B. Longitudinal nasopharyngeal carriage and antibiotic resistance of respiratory bacteria in indigenous Australian and Alaska native children with bronchiectasis. *PLoS One* **8,** e70478 (2013).
49. Descheemaeker, P. Macrolide resistance and erythromycin resistance determinants among Belgian *Streptococcus pyogenes* and *Streptococcus pneumoniae* isolates. *J Antimicrob Chemother* **45,** 167-173 (2000).
50. Varaldo, P.E., Montanari, M.P. & Giovanetti, E. Genetic elements responsible for erythromycin resistance in streptococci. *Antimicrob Agents Chemother* **53,** 343-53 (2009).
51. Stadler, C. & Teuber, M. The Macrolide Efflux Genetic Assembly of *Streptococcus pneumoniae* Is Present in Erythromycin-Resistant *Streptococcus salivarius*. *Antimicrob Agents and Chemother* **46,** 3690-3691 (2002).
52. Seral, C., Castillo, F.J., Rubio-Calvo, M.C., Fenoll, A., Garcia, C. & Gomez-Lus, R. Distribution of resistance genes tet(M), aph3'-III, catpC194 and the integrase gene of Tn1545 in clinical *Streptococcus pneumoniae* harbouring erm(B) and mef(A) genes in Spain. *J Antimicrob Chemother* **47,** 863-6 (2001).
53. Sutcliffe, J., Tait-Kamradt, A. & Wondrack, L. *Streptococcus pneumoniae* and *Streptococcus pyogenes* resistant to macrolides but sensitive to clindamycin: a common resistance pattern mediated by an efflux system. *Antimicrob Agents Chemother* **40,** 1817-24 (1996).
54. Daly, M.M., Doktor, S., Flamm, R. & Shortridge, D. Characterization and prevalence of MefA, MefE, and the associated msr(D) gene in *Streptococcus pneumoniae* clinical isolates. *J Clin Microbiol* **42,** 3570-4 (2004).
55. Tettelin, H., Masignani, V., Cieslewicz, M.J., Eisen, J.A., Peterson, S., Wessels, M.R., Paulsen, I.T., Nelson, K.E., Margarit, I., Read, T.D., Madoff, L.C., Wolf, A.M., Beanan, M.J., Brinkac, L.M., Daugherty, S.C., DeBoy, R.T., Durkin, A.S., Kolonay, J.F., Madupu, R., Lewis, M.R., Radune, D., Fedorova, N.B., Scanlan, D., Khouri, H., Mulligan, S., Carty, H.A., Cline, R.T., Van Aken, S.E., Gill, J., Scarselli, M., Mora, M., Iacobini, E.T., Brettoni, C., Galli, G., Mariani, M., Vegni, F., Maione, D., Rinaudo, D., Rappuoli, R., Telford, J.L., Kasper, D.L., Grandi, G. & Fraser, C.M. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **99,** 12391-6 (2002).
56. Permina, E.A., Mironov, A.A. & Gelfand, M.S. Damage-repair error-prone polymerases of eubacteria: association with mobile genome elements. *Gene* **293,** 133-40 (2002).

57. Tocci, N., Iannelli, F., Bidossi, A., Ciusa, M.L., Decorosi, F., Viti, C., Pozzi, G., Ricci, S. & Oggioni, M.R. Functional analysis of pneumococcal drug efflux pumps associates the MATE DinF transporter with quinolone susceptibility. *Antimicrob Agents Chemother* **57,** 248-53 (2013).
58. Beuning, P.J., Simon, S.M., Godoy, V.G., Jarosz, D.F. & Walker, G.C. Characterization of Escherichia coli Translesion Synthesis Polymerases and Their Accessory Factors. in *Methods in Enzymology*, Vol. Volume 408 318-340 (Academic Press, 2006).
59. Brueggemann, A.B., Coffman, S.L., Rhomberg, P., Huynh, H., Almer, L., Nilius, A., Flamm, R. & Doern, G.V. Fluoroquinolone Resistance in *Streptococcus pneumoniae* in United States since 1994-1995. *Antimicrob Agents and Chemother* **46,** 680-688 (2002).
60. Korzheva, N., Davies, T.A. & Goldschmidt, R. Novel Ser79Leu and Ser81Ile substitutions in the quinolone resistance-determining regions of ParC topoisomerase IV and GyrA DNA gyrase subunits from recent fluoroquinolone-resistant *Streptococcus pneumoniae* clinical isolates. *Antimicrob Agents Chemother* **49,** 2479-86 (2005).
61. Perichon, B., Tankovic, J. & Courvalin, P. Characterization of a mutation in the parE gene that confers fluoroquinolone resistance in *Streptococcus pneumoniae*. *Antimicrob Agents and Chemother* **41,** 1166-1167 (1997).
62. Gillespie, S.H., Voelker, L.L., Ambler, J.E., Traini, C. & Dickens, A. Fluoroquinolone resistance in *Streptococcus pneumoniae*: evidence that gyrA mutations arise at a lower rate and that mutation in gyrA or parC predisposes to further mutation. *Microb Drug Resist* **9,** 17-24 (2003).
63. Weigel, L.M., Anderson, G.J., Facklam, R.R. & Tenover, F.C. Genetic analyses of mutations contributing to fluoroquinolone resistance in clinical isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **45,** 3517-23 (2001).
64. Pletz, M.W., Shergill, A.P., McGee, L., Beall, B., Whitney, C.G., Klugman, K.P. & Active Bacterial Core Surveillance, T. Prevalence of first-step mutants among levofloxacin-susceptible invasive isolates of *Streptococcus pneumoniae* in the United States. *Antimicrob Agents Chemother* **50,** 1561-3 (2006).
65. Paulsen, I.T., Brown, M.H. & Skurray, R.A. Proton-dependent multidrug efflux systems. *Microbiol Rev* **60,** 575-608 (1996).
66. Gill, M.J., Brenwald, N.P. & Wise, R. Identification of an Efflux Pump Gene,pmrA, Associated with Fluoroquinolone Resistance in *Streptococcus pneumoniae*. *Antimicrob Agents and Chemother* **43,** 187-189 (1999).
67. Mayfield, J.A., Dehner, C.A. & DuBois, J.L. Recent advances in bacterial heme protein biochemistry. *Curr Opin Chem Biol* **15,** 260-6 (2011).
68. Romero-Espejel, M.E., Gonzalez-Lopez, M.A. & Olivares-Trejo Jde, J. *Streptococcus pneumoniae* requires iron for its viability and expresses two membrane proteins that bind haemoglobin and haem. *Metallomics* **5,** 384-9 (2013).
69. Anzaldi, L.L. & Skaar, E.P. Overcoming the heme paradox: heme toxicity and tolerance in bacterial pathogens. *Infect Immun* **78,** 4977-89 (2010).
70. Friedman, D.B., Stauff, D.L., Pishchany, G., Whitwell, C.W., Torres, V.J. & Skaar, E.P. *Staphylococcus aureus* redirects central metabolism to increase iron availability. *PLoS Pathog* **2,** e87 (2006).
71. Lechardeur, D., Cesselin, B., Liebl, U., Vos, M.H., Fernandez, A., Brun, C., Gruss, A. & Gaudu, P. Discovery of intracellular heme-binding protein HrtR, which controls heme efflux by the conserved HrtB-HrtA transporter in Lactococcus lactis. *J Biol Chem* **287,** 4752-8 (2012).

72. Torres, V.J., Stauff, D.L., Pishchany, G., Bezbradica, J.S., Gordy, L.E., Iturregui, J., Anderson, K.L., Dunman, P.M., Joyce, S. & Skaar, E.P. A *Staphylococcus aureus* regulatory system that responds to host heme and modulates virulence. *Cell Host Microbe* **1,** 109-19 (2007).
73. Sachla, A.J. & Eichenbaum, Z. The GAS PefCD exporter is a MDR system that confers resistance to heme and structurally diverse compounds. *BMC Microbiol* **16**, 68 (2016).
74. Gill, M.J., Brenwald, N.P. & Wise, R. Identification of an Efflux Pump Gene, pmrA, Associated with Fluoroquinolone Resistance in *Streptococcus pneumoniae*. *Antimicrobial Agents and Chemotherapy* **43,** 187-189 (1999).
75. Pletz, M.W., Michaylov, N., Schumacher, U., van der Linden, M., Duesberg, C.B., Fuehner, T., Klugman, K.P., Welte, T. & Makarewicz, O. Antihypertensives suppress the emergence of fluoroquinolone-resistant mutants in pneumococci: an *in vitro* study. *Int J Med Microbiol* **303,** 176-81 (2013).
76. McGee, L., McDougal, L., Zhou, J., Spratt, B.G., Tenover, F.C., George, R., Hakenbeck, R., Hryniewicz, W., Lefevre, J.C., Tomasz, A. & Klugman, K.P. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. *J Clin Microbiol* **39,** 2565-71 (2001).
77. Weigel, L.M., Anderson, G.J., Facklam, R.R. & Tenover, F.C. Genetic Analyses of Mutations Contributing to Fluoroquinolone Resistance in Clinical Isolates of *Streptococcus pneumoniae*. *Antimicrobial Agents and Chemotherapy* **45,** 3517-3523 (2001).
78. Tankovic, J., Perichon, B., Duval, J. & Courvalin, P. Contribution of mutations in gyrA and parC genes to fluoroquinolone resistance of mutants of *Streptococcus pneumoniae* obtained *in vivo* and *in vitro*. *Antimicrob Agents and Chemother* **40,** 2505-2510 (1996).
79. Balsalobre, L., Ferrandiz, M.J., Linares, J., Tubau, F. & de la Campa, A.G. Viridans group streptococci are donors in horizontal transfer of topoisomerase IV genes to *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **47,** 2072-81 (2003).
80. Pletz, M.W., McGee, L., Beall, B., Whitney, C.G. & Klugman, K.P. Interspecies recombination in type II topoisomerase genes is not a major cause of fluoroquinolone resistance in invasive *Streptococcus pneumoniae* isolates in the United States. *Antimicrob Agents Chemother* **49,** 779-80 (2005).
81. Canton, R., Morosini, M., Enright, M.C. & Morrissey, I. Worldwide incidence, molecular epidemiology and mutations implicated in fluoroquinolone-resistant *Streptococcus pneumoniae*: data from the global PROTEKT surveillance programme. *J Antimicrob Chemother* **52,** 944-52 (2003).
82. World Health Organization. Introduction to Drug Utilization Research. (Oslo, Norway, 2003).
83. Polk, R.E., Fox, C., Mahoney, A., Letcavage, J. & MacDougall, C. Measurement of adult antibacterial drug use in 130 US hospitals: comparison of defined daily dose and days of therapy. *Clin Infect Dis* **44,** 664-70 (2007).
84. Fairlie, T., Shapiro, D.J., Hersh, A.L. & Hicks, L.A. National trends in visit rates and antibiotic prescribing for adults with acute sinusitis. *Arch Intern Med* **172,** 1513-4 (2012).
85. Grijalva, C.G., Nuorti, J.P. & Griffin, M.R. Antibiotic prescription rates for acute respiratory tract infections in US ambulatory settings. *JAMA* **302,** 758-66 (2009).
86. Hersh, A.L., Shapiro, D.J., Pavia, A.T. & Shah, S.S. Antibiotic prescribing in ambulatory pediatrics in the United States. *Pediatrics* **128,** 1053-61 (2011).

87. Steinman, M.A., Gonzales, R., Linder, J.A. & Landefeld, C.S. Changing use of antibiotics in community-based outpatient practice, 1991–1999. *Annals of Internal Medicine* **138**, 525-533 (2003).
88. McCaig, L.F., Besser, R.E. & Hughes, J.M. Antimicrobial Drug Prescriptions in Ambulatory-Care Settings, United States, 1992-2000. *Emerging infectious diseases* **9**, 432-437 (2003).
89. Marks, L.R., Reddinger, R.M. & Hakansson, A.P. High levels of genetic recombination during nasopharyngeal carriage and biofilm formation in *Streptococcus pneumoniae*. *MBio* **3**(2012).
90. McCavit, T.L., Quinn, C.T., Techasaensiri, C. & Rogers, Z.R. Increase in invasive *Streptococcus pneumoniae* infections in children with sickle cell disease since pneumococcal conjugate vaccine licensure. *J Pediatr* **158**, 505-7 (2011).
91. Daw, N.C., Wilimas, J.A., Wang, W.C., Presbury, G.J., Joyner, R.E., Harris, S.C., Davis, Y., Chen, G. & Chesney, P.J. Nasopharyngeal carriage of penicillin-resistant *Streptococcus pneumoniae* in children with sickle cell disease. *Pediatrics* **99**, E7 (1997).
92. Gardner, S.N. & Hall, B.G. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One* **8**, e81760 (2013).
93. Croucher, N.J., Walker, D., Romero, P., Lennard, N., Paterson, G.K., Bason, N.C., Mitchell, A.M., Quail, M.A., Andrew, P.W., Parkhill, J., Bentley, S.D. & Mitchell, T.J. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*Spain23F ST81. *J Bacteriol* **191**, 1480-9 (2009).
94. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
95. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-1 (2010).
96. Sahl, J.W., Caporaso, J.G., Rasko, D.A. & Keim, P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ* **2**, e332 (2014).
97. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84 (2002).
98. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
99. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-8 (2013).
100. R Core Development Team. R: a language and environment for statistical computing. 3.2.2 edn (Foundation for Statistical Computing, Vienna, Austria, 2010).
101. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
102. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).

Chapter 6

Genetic microbial correlates of the clinical manifestation of invasive pneumococcal disease

Fredrick M. Mobegi
Amelieke Cremers
Stephen Bentley
Marien I. de Jonge
Sacha A. F. T. van Hijum
Aldert Zomer

Manuscript in preparation

Abstract

Invasive pneumococcal disease (IPD), defined as an infection verified by the isolation of *Streptococcus pneumoniae* from a normally sterile body site, such as the blood and cerebrospinal fluid, remains a major burden to global healthcare systems, causing approximately 25% of all preventable deaths in children under the age of 5 and more than 1.3 million infant deaths yearly. The pneumococcus carries various virulence factors that help to evade host defenses and persist during colonization. Although these factors are involved in colonization, they may also play a role in pneumococcal infectiveness and the subsequently clinical manifestation of IPD. Genome-wide association studies (GWAS) are undoubtedly an important tool for genetic analysis is humans, aiding the identification of various genetic determinants of inherited traits and diseases. However, the application of GWAS to identify genotypic variants responsible for important bacterial phenotypes has only been made possible recently by the significant advances in next generation genome sequencing. Various studies are now utilizing GWAS with diverse study designs that incorporate a range of genetic features in the core genome such as single nucleotide polymorphisms (SNPs) and small indels, or variations in the accessory or flexible genome such as presence and/or absence of a genes to correlate measurable bacterial phenotypes. Here, we report the use of GWAS to discover how genetic variation in the accessory genomes of 350 *S. pneumoniae* isolates correlate with various clinical contingencies of invasive pneumococcal disease.

Introduction

Streptococcus pneumoniae remains an important human commensal and an opportunistic bacterium pathogen. It is the most common cause of acute otitis media in children, and community-acquired pneumonia, meningitis, and bacteremia in children and adults, accounting for multiple morbidities and over 1.3 million childhood mortalities worldwide each year [1]. The burden of invasive pneumococcal disease (IPD) is particularly high in children under the age of 2 years, adults above 65 years old, and individuals with predisposing conditions such as asplenia, chronic medical conditions like diabetes, or immunosuppressive illnesses, particularly HIV and AIDS [2].

S. pneumoniae encodes various virulence factors that help promote infections and pathologies by various means such as adherence and invasion, evasion of host defenses and nutrient acquisition [3]. Its efficient competence machinery enables rapid recombination of foreign DNA into its genome [4, 5]. The horizontal transfer of DNA, either via recombination or via phage transduction begets extensive allelic variations in the core genome and a broad flexible or accessory genome, which underlie phenotypic differences in important traits such as virulence and antibiotic resistance. While most pneumococcal virulence factors have been characterized, their role in invasive pneumococcal disease remains underexplored.

Genome-wide association studies (GWAS) have over the past decade become the *de facto* approach to identify candidate genetic variations associated with complex inherited human diseases [6]. Despite being broached nearly a decade ago that tools and knowledge drawn from eukaryotic GWAS could be employed in bacteriology to investigate the genotypic etiology of relevant phenotypes [7], GWAS has limitedly been applied to the pneumococcus. The use of GWAS to identify genetic variants associated with important bacterial phenotypes has only been made possible recently by advances in bioinformatics and genome sequencing technologies [8]. Studies employ either alleles counting or homoplasy counting to compute the key association indicators [9, 10].

The use of GWAS to identify genetic variants associated with disease manifestation still remains unexplored. A recently published set of 350 draft pneumococcal genomes isolated from adults admitted with IPD in Nijmegen [11, 12], whose clinical metadata is available, provided an opportunity to examine the role of genetic diversity in the accessory genome on the clinical manifestation of IPD. These isolates cover the period before and after introduction of 7-valent pneumococcal conjugate vaccine (PCV7) in the national vaccination program. We performed a GWAS analysis on the accessory genome using PLINK [13] and corrected for effects of clonal population structure. We identified a number of genetic factors associated with clinical manifestation of invasive pneumococcal disease and discuss their potential mode of action in disease.

Table 1. A summary of clinical phenotypes included in the study.

Type	Phenotype	Cases	Controls	Missing	Notes
Binary phenotypes	Empyema	31	317	1	a
	Empyema and pneumonia	30	255	64	a
	Meningitis	30	318	1	a
	Neutrophils (dichotomous)	242	36	71	b
	Pleural effusion	105	130	114	c
	Pneumonia	285	63	1	a
	Renal disease	19	329	1	d
	Systemic Inflammatory Response Syndrome (SIRS) on Admission	273	38	38	c
	30 day mortality	37	308	4	c
	Cardiovascular disease	148	200	1	d
	Cancer	75	273	1	d
	Collection during influenza season	207	142	0	d
	Chronic Obstructive Pulmonary Disease (COPD)	76	272	1	d
	Cough	209	102	38	b
	Dicompensatio Cordis	41	307	1	d
	Diabetes Mellitus	60	288	1	d

Type	Phenotype	Category	Denoted	Count	Notes
Continuous phenotypes	Age	Data missing	-9	1	d
		0-6 years	0	7	
		7-55 years	1	88	
		56 years or older	2	253	
	Pneumonia Severity Index (PSI)	Missing	-9	81	c
		Outpatient	1	9	
		Short inpatient	2	96	
		Hospitalized	3	106	
		ICU	4	58	
	C- reactive proteins (CRP)	Missing	-9	2	b
		Normal (<10mg/L)	0	8	
		Low risk (10-100mg/L)	1	70	
		Average risk (200-300mg/L)	2	111	
		High risk (300-500mg/L)	3	117	
		Very high risk (>500mg/L)	4	41	
	Charlson comorbidity index*	Missing	-9	1	d
		Normal	0	36	
		Myocardial Infarction	1	22	
		Congestive Heart Failure	2	22	
		Peripheral Vascular Disease	3	47	
		Cerebrovascular Disease	4	62	
		Dementia	5	56	
		COPD	6	30	
		Connective Tissue Disease	7	39	
		Peptic Ulcer Disease	8	18	
		Diabetes Mellitus	9	10	
		Kidney Disease	10	5	
		Liver Disease	15	1	

Notes: a, type of infection; b, disease characteristics; c, severity of disease; and d, risk factors; *http://touchcalc.com/calculators/cci_js.

Results and discussion

Defining the sequence clusters and the accessory genome

Coding sequences from the 350 isolates have been grouped into a total of 3,021 clusters of orthologous genes (OGs), of which 1,075 were core; appearing in a single copy in all isolates and 1,946 accessory OGs, containing rarer or variable genotypes [12]. Filtering for OGs containing singletons and genes present in less than 2% (7 isolates) of the isolates resulted into 1,532 accessory OGs. Analysis of the concatenated variant sites on the core OGs using hierarchical BAPS [14] converged at 13 sequence clusters (SCs); 12 largely monophyletic SCs and a thirteenth comprised of atypical genotypes (Figure 1). After collating measures according to the type of pneumococcal disease, severity of disease, disease characteristics, and underlying risk factors, a total of 16 binary phenotypes and 4 multi-category phenotypes were included in the analysis (Table 1).

Although disease is a highly polygenic phenomenon, disease phenotypes show positive associations with certain SCs (Fishers *p*-value <0.05 Bonferroni corrected for multiple testing), indicating that specific clones are likely to cause more disease that others (Table 2), or that different genetic signatures result in the same disease phenotype. Therefore, corrections for population structure stratifications is required. For example SC1, SC6, and SC10 correlate with pneumococcal disease. These SCs respectively consist of serotypes 14, 8, and 1, which have been observed to be the dominant serotypes that cause invasive pneumococcal infections [12, 15]. Surprisingly, rarer sequence types (SC2; consisting of atypical genotypes) also seem to cause more disease. Studies have shown that rarer pneumococcal serotypes might be more prevalent in disease than previously thought [16]. SNPs analysis may be necessary to elucidate if specific polymorphisms in the clones underlie these associations. Highly recombinogenic bacteria like the pneumococcus exhibit clonal population structures. The linkage disequilibrium (LD) between gene presence or absence and phylogenetically informative SNPs. Unfortunately, we lack sufficient number of genomes with clinical metadata to perform the analyses on the atypical genotypes.

The Cochran–Mantel–Haenszel statistic was employed to infer associations between pneumococcal accessory genome and binary phenotypes with sequence cluster-membership as covariate. The results in Supplementary Table 1 (stratified) and Supplementary Table 2 (unstratified) show that correcting for population substructure stratification minimized the inflation of association *p-values*. Correction was specific to variation in frequency of candidate OGs across ancestral populations. Without stratification, there was a substantial systematic inflation of the association test *p*-values across all clinical phenotypes. For example, OG 2136 and OG 2380 show significant association with empyema (*p*-value <0.05) before stratification but not so after population stratification. Therefore, stratification minimizes spurious associations and

maximizes the power to detect true associations in GWAS studies. Only three phenotypes (pneumoniae, meningitis, and 30 day mortality) exhibited significant associations (*p*-values < 0.05 Bonferroni adjusted for multiple testing) after correcting for population stratification (Figure 1; Table 3).

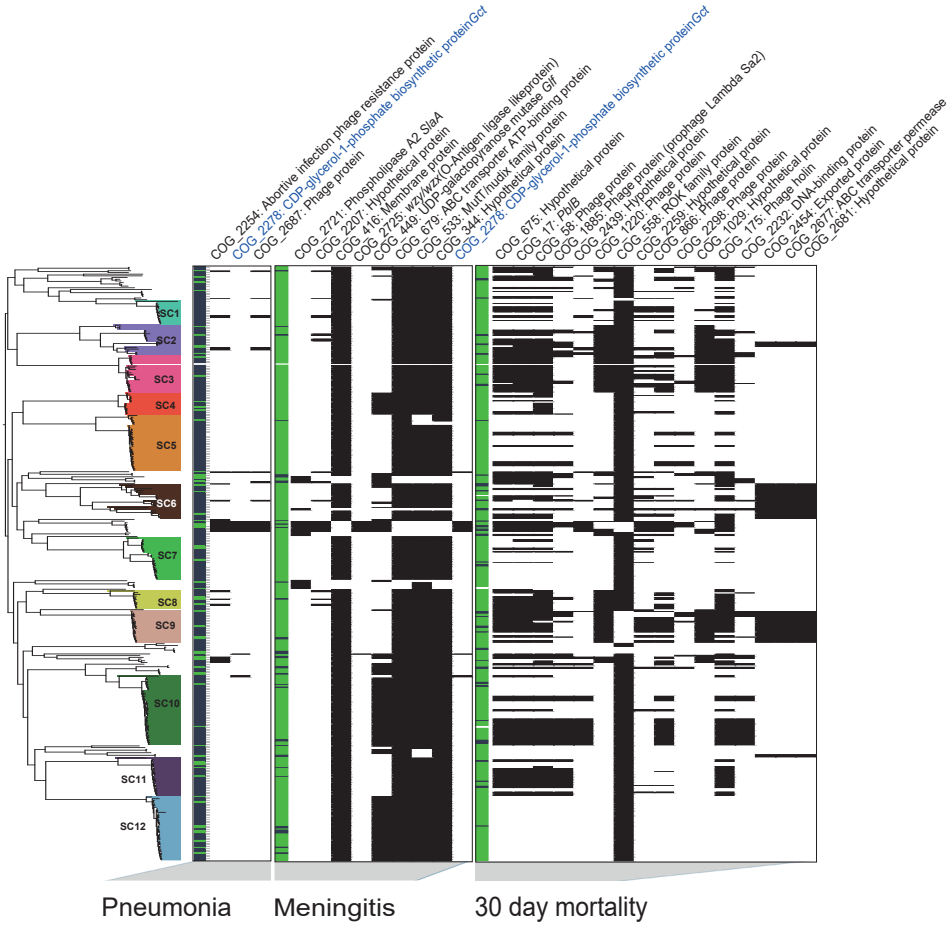


Figure 1. A phylogeny of the variable sites from the core genome of all clinical isolates used in study. Clades are colored according to the sequence clusters as determined using BAPS (see materials and methods). Presence (black) or absence (white) of genes statistically associated with pneumonia, meningitis and 30-day mortality is represented by the heatmap. The first column of each disease eventuality shows the phenotype for individual isolates; green=control, black = case.

Table 2. Statistical association between sequence clusters and phenotypes of disease manifestation

Sequence cluster	Meningitis +	Meningitis -	Others diseases+	Others diseases-	*p*-value
SC1	1	18	29	302	1
SC2	15	57	15	263	0.000155556
SC3	0	25	30	295	0.1487862
SC4	0	13	30	307	0.6133904
SC5	2	21	28	299	1
SC6	3	38	27	282	1
SC7	0	22	30	298	0.2380451
SC8	0	15	30	305	0.6274878
SC9	5	33	25	287	0.3502867
SC10	1	32	29	288	0.3363776
SC11	1	10	29	310	1
SC12	2	16	28	304	0.6596066
SC13	0	20	30	300	0.238468
	Pneumonia +	**Pneumonia -**	**Others diseases+**	**Others diseases-**	***p*-value**
SC1	15	4	270	61	0.7630057
SC2	48	24	237	41	0.000606396
SC3	22	3	263	62	0.5926162
SC4	10	3	275	62	0.7150978
SC5	19	4	266	61	1
SC6	39	2	246	63	0.01691588
SC7	19	3	266	62	0.7774798
SC8	14	1	271	64	0.3212966
SC9	29	9	256	56	0.3814105
SC10	32	1	253	64	0.01615192
SC11	8	3	277	62	0.4351478
SC12	13	5	272	60	0.3468605
SC13	17	3	268	62	1
	30-day mortality +	**30-day mortality -**	**Others diseases+**	**Others diseases-**	***p*-value**
SC1	5	14	23	308	0.0120289
SC2	2	70	26	252	0.08613299
SC3	2	23	26	299	1
SC4	0	13	28	309	0.610641
SC5	3	20	25	302	0.4132743
SC6	5	36	23	286	0.3514003
SC7	2	20	26	302	0.6921759
SC8	0	15	28	307	0.6199427
SC9	1	37	27	285	0.3385618
SC10	1	32	27	290	0.4962165
SC11	0	11	28	311	1
SC12	3	15	25	307	0.1658858
SC13	4	16	24	306	0.06489312

Genes associated with pneumonia and meningitis

Certain proteins or enzymes expressed by the pneumococcus significantly contribute to pathogenesis and might be involved in the disease process caused by these pathogen. Indeed, several pneumococcal virulence factors including choline binding protein A (CbpA) [17], pneumococcal surface antigen A (PsaA) [18], pneumococcal surface protein A (PspA) [19], hyaluronate lyase [20], pneumolysin [21], neuraminidases [20], and major autolysin (LytA) [20], and their role in pneumococcal pathogenesis have been described. These proteins often interact directly with the host tissues and may also be involved in camouflaging the pneumococcal surface from the host defenses.

The capsule polysaccharide (CPS) plays an important role in the virulence of *S. pneumoniae* and provides resistance to phagocytosis. Presence of 'CDP-glycerol-1-phosphate biosynthetic protein' (OG_2278; *gct*), a gene involved in the biosynthesis of lipoteichoic acid precursors of the CPS [22], is underrepresented in pneumonia and overrepresented in meningitis; *p*-values 0.0098 and 0.001 respectively, Bonferroni corrected (Table 3). Teichoic acids (TAs) mainly provide rigidity to the cell-wall by attracting cations such as magnesium and sodium [23]. Bacterial strains in which expression TAs synthesis has been prevented are biologically nonviable and show profound morphological aberrations. In addition, TAs may aid also in regulating cell growth by limiting autolysins from breaching the β (1-4) bond between N-acetylmuramic acid and N-acetylglucosamine. Since the bacteria exhibiting increased cell wall rigidity could better resist killing by the host immune defenses, we hypothesize that the TAs may play a role in persisting the host protection during invasive infection. Although not conclusively supported, studies have also implicated TAs in biofilm formation, host tissue adhesion, cell growth, division and morphogenesis, and as receptor molecules for some Gram-positive bacteriophages [24-26].

The abortive infection phage resistance protein (OG_2254) and an uncharacterized phage protein (OG_2287; last gene in phage lytic cycle, downstream the phage transcription regulator-see figure 3) are overrepresented in isolates that caused pneumonia compared to isolates that caused other type of infections. Phages are perhaps the most diversified, adaptive and widespread microorganisms [27]. The abortive infection (Abi) systems protects the host bacterium against an existing phage infection by promoting cell death and limit phage replication [28] and might have additional roles like enabling stabilization of integrative and conjugative elements [29], some of which encode virulence and resistance genes that are transmissible between Streptococci [30, 31]. The direct role of these elements in promoting pneumonia is not clear but phages may carry virulence factors which we assume may enhance pathogenesis [32, 33]

.

Phenotype	OG ID	Annotation	Isolates with OG	% of isolates with phenotype carrying the OG	% of isolates without phenotype carrying the OG	p-value	Adjusted p-value	#Isolates before vaccine	#Isolates after vaccine	% prevelence before vaccine	% prevalence after vaccine
	558	Putative Rok Family Protein	318	24,32%	71,43%	0,00254	0,00223	165	153	88,71%	93,87%
	58	Phage Protein	178	18,92%	52,92%	8,83E-05	0,00102	104	74	55,91%	45,40%
	175	Ll-H Family Phage Holin	172	81,08%	45,13%	3,76E-05	0,00433	100	72	53,76%	44,17%
	17	PblB	167	72,97%	44,48%	0,00144	0,00034	98	69	52,69%	42,33%
	675	Hypothetical Protein	164	75,68%	43,18%	0,000198	0,00023	96	68	51,61%	41,72%
	866	Phage Protein	106	51,35%	27,60%	0,004306	0,00359	65	41	34,95%	25,15%
	1220	Phage Protein	94	48,65%	24,35%	0,00291	0,00203	65	29	34,95%	17,79%
30 day mortality	1885	Phage Protein, Prophage Lambdasa2 Single-Strand Bifunctional Prote	85	45,95%	21,43%	0,001954	0,00192	47	38	25,27%	23,31%
	1029	Hypothetical Protein	79	43,24%	20,13%	0,003037	0,00374	57	22	30,65%	13,50%
	2259	Hypothetical Protein	64	37,84%	15,91%	0,002678	0,00276	41	23	22,04%	14,11%
	2454	Exported Protein	45	29,73%	10,71%	0,002964	0,05296	25	20	13,44%	12,27%
	2677	ABC Transporter Permease	45	29,73%	10,71%	0,002964	0,05431	25	20	13,44%	12,27%
	2232	DNA-Binding Protein,Phage Membrane Protein	41	27,03%	9,74%	0,004869	0,04523	24	17	12,90%	10,43%
	2439	Hypothetical Protein	41	29,73%	9,42%	0,001226	0,00201	25	16	13,44%	9,82%
	2298	Phage Protein	27	21,62%	6,17%	0,004033	0,00369	17	10	9,14%	6,13%
	1144	Hypothetical Protein	347	6,67%	0,00%	0,007205	0,09994	186	161	100,00%	98,77%
	416	Membrane Protein	330	20,00%	4,09%	0,003028	0,00073	172	158	92,47%	96,93%
	679	ABC Transporter Atp-Binding Protein	330	20,00%	4,09%	0,003028	0,00095	172	158	92,47%	96,93%
	344	Hypothetical Protein	328	20,00%	4,72%	0,005339	0,001	170	158	91,40%	96,93%
	533	Mutt/Nudix Family Protein	323	23,33%	5,98%	0,003594	0,00097	170	153	91,40%	93,87%
	1521	Hypothetical Protein	297	33,33%	13,21%	0,006623	0,09989	162	135	87,10%	82,82%
	449	UDP-Galactopyranose Mutase Glf	130	63,33%	34,91%	0,002866	0,00095	60	70	32,26%	42,94%
Meningitis	2034	Hypothetical Protein	54	0,00%	16,98%	0,007421	0,09995	25	29	13,44%	17,79%
	1616	Ci-Like Repressor	22	20,00%	5,03%	0,006893	0,09991	15	7	8,06%	4,29%
	2207	Hypothetical Protein	21	23,33%	4,40%	0,000868	0,00059	16	5	8,60%	3,07%
	2721	Phospholipase A2 Slaa	19	20,00%	4,09%	0,003028	0,00045	14	5	7,53%	3,07%
	2687	Phage Protein	16	16,67%	3,46%	0,007547	0,09996	10	6	5,38%	3,68%
	2278	CDP-Glycerol-1-Phosphate Biosynthetic Protein Gct	10	13,33%	1,89%	0,006593	0,00999	5	5	2,69%	3,07%
	2725	Wzy/Wzx; O-Antigen Ligase-Like Protein	8	13,33%	1,26%	0,002489	0,00092	5	3	2,69%	1,84%
	2992	Hypothetical Protein	2	6,67%	0,00%	0,007205	0,09994	1	1	0,54%	0,61%
	2601	Sodium:Solute Symporter Family Protein	42	14,74%	0,00%	0,000178	0,1834	20	22	10,75%	13,50%
Pneumonia	2254	Abortive Infection Phage Resistance Protein	24	4,56%	17,46%	0,001074	0,00705	16	8	8,60%	4,91%
	2687	Phage Protein	16	2,81%	12,70%	0,002922	0,00965	10	6	5,38%	3,68%
	2278	CDP-Glycerol-1-Phosphate Biosynthetic Protein Gct	10	1,40%	9,52%	0,003264	0,00976	5	5	2,69%	3,07%

Table 3. A summary of the genes statistically associated with disease.

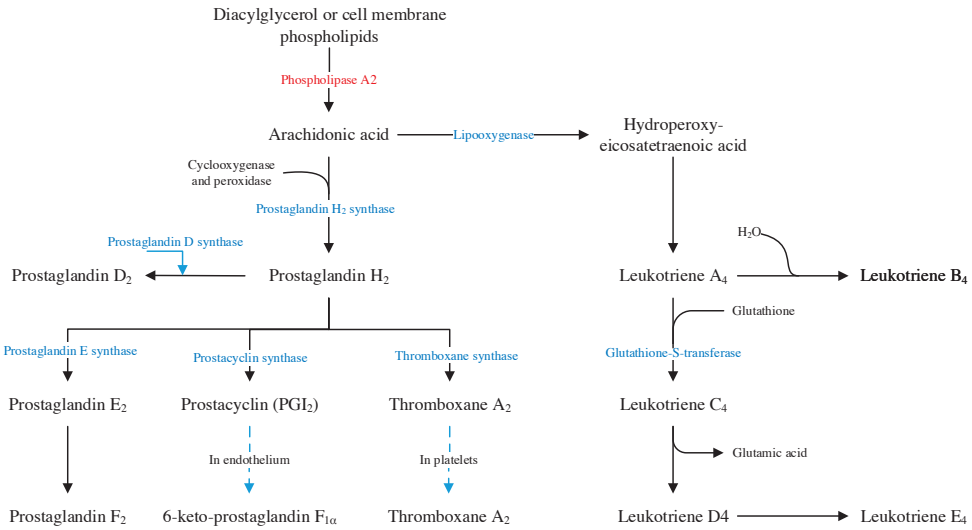


Figure 2. A schematic overview of eicosanoids biosynthesis

We identified some proteins, which significantly associated with meningitis, and have been previously associated with increased virulence and observed in meningitis. They include phospholipase A2; *slaA* [34], *wzy*/*wzx* [22], and UDP-galactopyranose mutase; *glf* [35]. Phospholipase A2, encoded by *slaA*, has a function in host-pathogen interaction as a colonization and pathogenesis promoting factor [36]. *In vitro* studies have shown that SlaA promotes colonization by cleaving epithelial cell membrane phospholipids in the nasopharynx releasing arachidonic acid [37]. Current knowledge on meningeal infection by pneumococcus implicates platelet-activating factor receptor (PAFR) in binding *S. pneumoniae*, thereby aiding adhesion, uptake and transcytosis through endothelial cells [38, 39]. Pneumococcus interacts with PAFR which activates signal transduction pathways including phospholipase A2 [40] to release arachidonic acid [41]; a precursor of eicosanoids including prostaglandins and leukotrienes (Figure 2). Also, during meningeal infection by Group B Streptococci, the host cytosolic phospholipase $A_2\alpha$ facilitates the release of arachidonic acid from brain endothelial membrane [34]. Arachidonic acid is processed into leukotrienes that regulate permeability of the human brain microvascular endothelial cells (HBMEC), which constitute the blood-brain barrier. Penetration of the blood-brain barrier is the crucial stage necessary for the development of pneumococcal meningitis. It is tempting to speculate that the bacterial SlaA from the pneumococcus may also releases arachidonic acid from the brain endothelial membrane, subsequently increasing permeability of the blood-brain barrier through leukotriene function. SlaA may therefore be a colonization factor that also has an inadvertent function of magnifying invasive pneumococcal infections by enabling the microorganism to cross the blood-brain barrier. *SlaA* together with streptococcal pyrogenic exotoxin K (*SpeK*), are implicated in

increased virulence, particularly in the serotype M3 clone of Group A Streptococcus (GAS) [42, 43]. Interestingly in our genomes, *SlaA* is neighbored by *SpeK*, perhaps an indicator of their function or relatedness as bacteriophage-encoded factors. However, these hypotheses have not been experimentally confirmed. An arachidonic acid release assay using HBMEC cells exposed to SlaA expressing pneumococcus would be of great benefit. An additional meningitis mouse model with pneumococcal strains expressing SlaA and isogenic knockouts could be used to conclusively confirm the role of SlaA in meningitis.

Apart from pneumococcal serotypes 3 and 37 that are synthesized by the synthase pathway [41, 44-46]; other pneumococcal capsular polysaccharides are generally synthesized by the Wzx/Wzy-dependent pathway [22]. The Wzy polymerase (*wzy*), Wzx flippase (*wzx*) are localized in the pneumococcal capsular biosynthetic (*cps*) loci along with various enzymes that modify the repeat units or add other moieties on capsule. The *wzy*/*wzx* genes are involved in polymerization and export of the pneumococcal capsule and are unique to each capsular type [22, 47]. The *wzy*/*wzx* genes (OG 2725) associated with meningitis in the Nijmegen adults' cohort are all unique to serotypes 18B/C. Serotype 18C is a highly invasive serotype [48] that has been included in all PCVs owing to its high global prevalence [49]. Overall, serogroup 18 has been reported to be among the most common serogroups causing meningitis which could explain the observed correlation [50, 51].

UDP-galactopyranose (OG 449) is an important constituent of glycoconjugates that comprise major portions of the cell surface and plays a crucial role in the infectious cycle [35]. OG 449 codes for UDP-galactopyranose which plays a role in production of vaccine-type capsules 6C, 7F, 18C, and non-vaccine capsules 6A, 8, 9V, 18B, 19A, 22A/F, 33A/F and 35B. Increased post-vaccine prevalence of vaccine related serotypes like 6A and 19A has been reported both in carriage [52] and invasive disease [12]. In addition to 7F and 8 were the dominant causes of post-vaccine invasive disease [12], while other non-vaccine serotypes including 11, 12, 15, 22F, 23A, 33F, 24, 34, and 35B are rapidly increasing in prevalence worldwide [53-57]. This increased prevalence and persistence may correlate to the increased propensity of these serotypes to infect and progress to meningitis. Serotype 35B has particularly been reported in infant pneumococcal meningitis [53] perhaps indicative of the increased invasiveness of this serotype.

Genes associated with 30 days mortality

Some IPD patients do not survive the first 30 days of hospitalization; herein referred to as 30-day mortality. In our dataset, the 30-day mortality phenotype was predominantly associated with bacteriophage genes (Figure 1). This association was primarily with genes at the lytic end of the phage (the late region) but also included other genes in the phage lytic cluster (Figure 3). Bacteriophages are among the most common and diverse entities in the atmosphere. They encode as few as four genes, to as many as hundreds of genes.

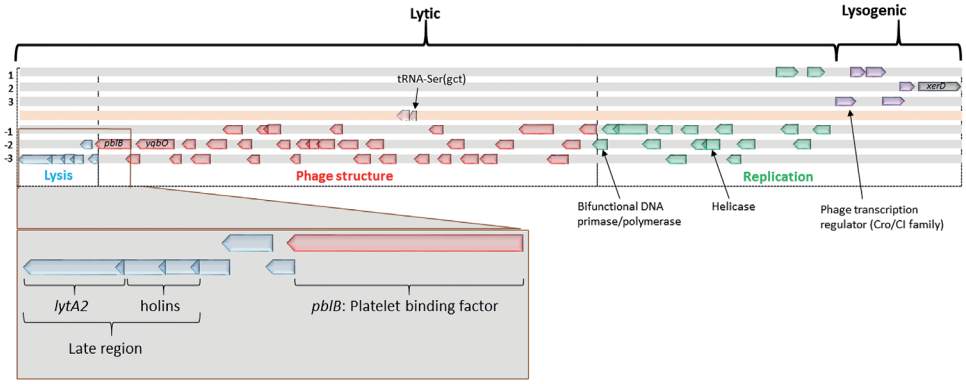


Figure 3. Lytic and lysogenic clusters of the Spn1 phage. The organization of phage genes was sketched from visualization of a whole-genome of an isolate carrying the *pblB* gene.

Bacteria are generally believed to acquire virulence properties from bacteriophages [58, 59]. Phage-encoded proteins have previously been implicated in providing a large gene pool for the pneumococcus which facilitate colonization and virulence [33, 60, 61].

Jeffrey N. Weiser and colleagues published this bacteriophage (named 'Spn1') containing 63 open reading frames in the pneumococcus [60]. They investigated the effect of Spn1 phage on colonization *in vivo* and observed that its presence was associated with reduced fitness during competitive colonization. Compared to the strain without the phage, the strain carrying Spn1 exhibited a defect in LytA-mediated autolysis *in vitro* but were characterized by an increased chain length of the cell wall, and increased resistance to lysis by penicillin. Expression of the phage genes was induced upon exposure to mitomycin C, but the treatment did not trigger lysis of the bacterial host [60].

The late phage region is expressed as the final step when the phage is induced. Induction may be caused by damage of the bacterial host DNA due to various factors, including chemical or oxidative compounds, UV irradiation, and halted translation by some antibiotics. To maintain the integrity of their genomes, bacteria turns on the DNA repair pathways which is governed by the so called 'SOS response' [60, 62]. The final step in phage replication consists of expression the late region which contains lytic enzymes that lyse the bacterial host, thereby releasing newly formed phage particles. However, when the bacterial host cells are compromised by the phage replication mediated lysis, pneumolysin toxin may also be released. Pneumolysin is a virulence factor that promotes adherence during colonization [63], damages host cells and interferes with the host immune response during infection [64]. It has no secretion system, but relies on lysis of bacterial host to be released in its environment. Therefore, rapid release of pneumolysin due to phage mediated lysis may be partly responsible for the increased mortality

observed in the patients. An initial mitomycin-C phage induction experiment did not result in phage plaque formation, thereby casting doubt on the lysis hypothesis, similar to what has been described for Spn1 [60].

We also observed that presence of *pblB*: a phage-encoded gene is associated with 30-day mortality. The *pblB* gene is also encoded in the late region of the same phage discussed above. PblB has been reported to promote platelet binding in *Streptococcus mitis* [65, 66]. In pneumonia infection, PblB has been reported to significantly enhance the platelet activating properties of pneumococci [33]. Higher platelet activation has been associated with myocardial infarction in pneumonia and the formation of microthrombi in other diseases, which may explain the potential of pneumococci carrying genes encoding PblB to cause endocarditis, cardiac failure and multi-organ failure, which could lead to death. Weiser and colleagues recently reported decreased bacterial fitness and increased resistance to penicillin by pneumococci carrying a phage encoding *PblB* [60]. Weiser *et al.* suggest that due to the use of penicillin, there was a positive selection on carriage of the phage, even though it has a fitness effect in carriage [60]. Inadvertently, this may have resulted in increased carriage of the pathogenicity factors encoded on the phages. To experimentally verify if the presence of the phage is associated with higher platelet activation, platelet-rich plasma (PRP) and whole blood can be used for *in vitro* stimulation assays with strains of the pneumococcus carrying or not carrying the phage. Flow cytometry can then be used to measure platelet activation and platelet-monocyte complex formation following induction of the phage.

Effect of vaccination on prevalence of disease associated genes

Significant post-vaccine restructuring of the pneumococcal population has been observed in both carriage [52, 67] and invasive disease [12, 52, 68]. The reshuffling is characterized by an initial decrease of a number of different genes within isolates followed by a return to equilibrium with a similar diversity of genes as in pre-vaccination state. Non-vaccine type isolates have become the dominant pneumococcal species after vaccination [12, 52, 68], virulence related genes may be lost or isolates with specific virulence factors may have an increased presence following vaccination [11, 69]. It is still unclear whether these perturbations affect virulence factors and consequently the eventuality of pneumococcal disease. With the exception of OGs 344, 416, 449, 533, 558, 679, 2034, 2278, 2292 and 2601 which increased in percentage prevalence, we observed a decrease in the prevalence of genes associated with disease (Table 3). A possible explanations is that these factors decrease with vaccine-types upon immunization but return with the replacing non-vaccine serotypes afterwards. Indeed, post-vaccine re-expansion of the genome content with genes circulating in population pre-vaccine has been reported [12, 52].

In conclusion, we suggest the use of whole genome sequencing for investigating the effect of vaccination on the prevalence of disease inflating factors in circulating

pneumococcal strains. This could give a forecast on the course of disease and facilitate modelling of future interventions. Additional experiments will be required to corroborate the observed statistical associations and possibly characterize the underlying biological mechanisms.

Methods

Determining the core and accessory genome

The isolates and clusters of orthologous genes (OGs) used in this analysis were defined previously [11, 12]. All putative protein coding sequences (CDSs) were analyzed in an all-versus-all protein blast using an *e-value* cut-off of 10e-15 and a BLOSUM90 substitution matrix. Orthologs were then grouped using TribeMCL [70] by implementing the Markov Clustering (MCL) step with an inflation value of 4. In total, the CDSs were grouped into 3,021 OGs, 1,075 of which were denominated as 'core' as they consisted of proteins present in a single copy in each of the 350 isolates. The remaining OGs were denominated as the flexible or accessory genome.

Defining the sequence clusters (SCs)

Protein coding sequences of the core OGs were aligned with MUSCLE [71] using a the default maximum of tree-dependent refinement iterations (16) for realignment until convergence was reached. The aligned sequences were then codon translated into nucleotide alignments using RevTrans [72]. Sequences in OGs encoding ribosomal proteins were then concatenated into a single 'ribosomal' alignment. A reference maximum likelihood phylogeny was constructed on the ribosomal alignment using RAxML v8.2.0 [73]. Phylogenies for each of the remaining core OGs were also constructed independently. The Euclidian distances (EuD) of these individual phylogenies were plotted against the ribosomal phylogeny resulting into three distinct distributions (Supplementary 3). To improve the core phylogeny resolution, sequences in OGs whose phylogenies were similar to the ribosomal phylogeny (EuD ≤ 0.03) were concatenated to the ribosomal alignment to give a single reduced 'core' super-alignment.

Finally, a maximum likelihood phylogenetic tree was constructed using RAxML and an alignment of all the concatenated polymorphic sites from 'core' super-alignment. The maximum-likelihood ratios were calculated using the general time-reversible model with a γ correction for site variation as the nucleotide substitution model. The support for nodes on the trees were confirmed using 100 random bootstrap replicates. Resulting phylogenetic trees were visualized in iTOL v3.0 [74]. Sequence clusters (SCs) were defined by analyzing the core super-alignment using hierarchical BAPS software v6.0 [14]. Two runs of 40 and 50 maximum clusters were performed and the consensus was established as previously described [12].

Testing for associations between presence or absence of a gene and clinical phenotypes

OGs representing the accessory genome were assigned a binary measure of presence (1) or absence (0) of a representative CDS from each strain. Applying this criterion, the contribution of each strain to the OGs was therefore denoted by a single numeric value (1 or 0) to represent the presence or absence of a gene, or a group of paralogs in a genome accordingly. Clinical phenotypes were classified into two types: binary or dichotomous phenotypes containing either cases (clinical outcome present) or controls (clinical outcome absent) for members within the cohort respectively, and multi-category ordinal phenotypes containing continuous or class-series traits like age and disease severity or comorbidity indices.

The Cochran–Mantel–Haenszel statistic, implemented in PLINK [13], was employed to infer associations between pneumococcal accessory genome and binary phenotypes. Associations for multi-category (non-binary) phenotypes like age, pneumonia severity index (PSI), and C-reactive proteins (CRP) were performed using multinomial logistic, and proportional odds ordinal logistic analyses implemented in Trinculo v0.9 (http://sourceforge.net/projects/trinculo/). All association tests were performed contingent on the SCs (representing genetic subpopulations within the population) as covariates of population substructure stratification. OGs showing statistically robust associations (p-values < 0.05 Bonferroni adjusted for multiple testing) were selected as putative causative candidates.

Supporting data

Supplementary material for this chapter are available online at:

https://www.dropbox.com/sh/bnx7vad8qivax1d/AABXGROkQDjovRQqmRruQqPOa?dl=0

References

1. Walker, C.L., Rudan, I., Liu, L., Nair, H., Theodoratou, E., Bhutta, Z.A., O'Brien, K.L., Campbell, H. & Black, R.E. Global burden of childhood pneumonia and diarrhoea. *Lancet* **381,** 1405-16 (2013).
2. Naheed, A., Saha, S.K., Breiman, R.F., Khatun, F., Brooks, W.A., El Arifeen, S., Sack, D., Luby, S.P. & Group, P.S. Multihospital Surveillance of Pneumonia Burden among Children Aged <5 Years Hospitalized for Pneumonia in Bangladesh. *Clinical Infectious Diseases* **48,** S82-S89 (2009).
3. Kadioglu, A., Weiser, J.N., Paton, J.C. & Andrew, P.W. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat Rev Micro* **6,** 288-301 (2008).
4. Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P. & Bentley, S.D. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46,** 305-9 (2014).
5. Croucher, N.J., Coupland, P.G., Stevenson, A.E., Callendrello, A., Bentley, S.D. & Hanage, W.P. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* **5,** 5471 (2014).
6. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M. & Wray, N.R. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44,** 247-250 (2012).
7. Falush, D. & Bowden, R. Genome-wide association mapping in bacteria? *Trends Microbiol* **14,** 353-5 (2006).
8. Read, T.D. & Massey, R.C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Medicine* **6,** 109 (2014).
9. Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E., Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., Turner, P., Bentley, S.D. & Parkhill, J. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet* **10,** e1004547 (2014).
10. Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., Bentley, S.D., Maiden, M.C.J., Parkhill, J. & Falush, D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the National Academy of Sciences* **110,** 11923-11927 (2013).
11. Cremers, A.J., Meis, J.F., Walraven, G., Jongh, C.E., Ferwerda, G. & Hermans, P.W. Effects of 7-valent pneumococcal conjugate 1 vaccine on the severity of adult 2 bacteremic pneumococcal pneumonia. *Vaccine* **32,** 3989-94 (2014).
12. Cremers, A.J.H., Mobegi, F.M., de Jonge, M.I., van Hijum, S.A.F.T., Meis, J.F., Hermans, P.W.M., Ferwerda, G., Bentley, S.D. & Zomer, A.L. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. *Scientific Reports* **5,** 14952 (2015).
13. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559-75 (2007).

14. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30,** 1224-8 (2013).
15. Ding, F., Tang, P., Hsu, M.-H., Cui, P., Hu, S., Yu, J. & Chiu, C.-H. Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14. *BMC Genomics* **10,** 1-13 (2009).
16. Sherwin, R.L., Gray, S., Alexander, R., McGovern, P.C., Graepel, J., Pride, M.W., Purdy, J., Paradiso, P. & File, T.M., Jr. Distribution of 13-valent pneumococcal conjugate vaccine *Streptococcus pneumoniae* serotypes in US adults aged >/=50 years with community-acquired pneumonia. *J Infect Dis* **208,** 1813-20 (2013).
17. Rosenow, C., Ryan, P., Weiser, J.N., Johnson, S., Fontan, P., Ortqvist, A. & Masure, H.R. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*. *Molecular Microbiology* **25,** 819-829 (1997).
18. Sampson, J.S., O'Connor, S.P., Stinson, A.R., Tharpe, J.A. & Russell, H. Cloning and nucleotide sequence analysis of psaA, the *Streptococcus pneumoniae* gene encoding a 37-kilodalton protein homologous to previously reported *Streptococcus* sp. adhesins. *Infect Immun* **62,** 319-24 (1994).
19. McDaniel, L.S., Sheffield, J.S., Delucchi, P. & Briles, D.E. PspA, a surface protein of *Streptococcus pneumoniae*, is capable of eliciting protection against pneumococci of more than one capsular type. *Infection and Immunity* **59,** 222-228 (1991).
20. Lock, R.A., Paton, J.C. & Hansman, D. Purification and immunological characterization of neuraminidase produced by *Streptococcus pneumoniae*. *Microbial Pathogenesis* **4,** 33-43 (1988).
21. Feldman, C., Munro, N.C., Jeffery, P.K., Mitchell, T.J., Andrew, P.W., Boulnois, G.J., Guerreiro, D., Rohde, J.A., Todd, H.C., Cole, P.J. & et al. Pneumolysin induces the salient histologic features of pneumococcal infection in the rat lung *in vivo*. *Am J Respir Cell Mol Biol* **5,** 416-23 (1991).
22. Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrell, B., Reeves, P.R., Parkhill, J. & Spratt, B.G. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* **2,** e31 (2006).
23. Swoboda, J.G., Campbell, J., Meredith, T.C. & Walker, S. Wall Teichoic Acid Function, Biosynthesis, and Inhibition. *Chembiochem : a European journal of chemical biology* **11,** 35-45 (2010).
24. Raisanen, L., Draing, C., Pfitzenmaier, M., Schubert, K., Jaakonsaari, T., von Aulock, S., Hartung, T. & Alatossava, T. Molecular interaction between lipoteichoic acids and Lactobacillus delbrueckii phages depends on D-alanyl and alpha-glucose substitution of poly(glycerophosphate) backbones. *J Bacteriol* **189,** 4135-40 (2007).
25. Schirner, K., Marles-Wright, J., Lewis, R.J. & Errington, J. Distinct and essential morphogenic functions for wall- and lipo-teichoic acids in Bacillus subtilis. *The EMBO Journal* **28,** 830-842 (2009).
26. Vergara-Irigaray, M., Maira-Litrán, T., Merino, N., Pier, G.B., Penadés, J.R. & Lasa, I. Wall teichoic acids are dispensable for anchoring the PNAG exopolysaccharide to the *Staphylococcus aureus* cell surface. *Microbiology (Reading, England)* **154,** 865-877 (2008).
27. Labrie, S.J., Samson, J.E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat Rev Micro* **8,** 317-327 (2010).

28. Fineran, P.C., Blower, T.R., Foulds, I.J., Humphreys, D.P., Lilley, K.S. & Salmond, G.P.C. The phage abortive infection system, ToxIN, functions as a protein–RNA toxin–antitoxin pair. *Proceedings of the National Academy of Sciences* **106,** 894-899 (2009).
29. Dy, R.L., Przybilski, R., Semeijn, K., Salmond, G.P.C. & Fineran, P.C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Research* (2014).
30. Haenni, M., Saras, E., Bertin, S., Leblond, P., Madec, J.-Y. & Payot, S. Diversity and Mobility of Integrative and Conjugative Elements in Bovine Isolates of *Streptococcus agalactiae*, S. dysgalactiae subsp. dysgalactiae, and S. uberis. *Applied and Environmental Microbiology* **76,** 7957-7965 (2010).
31. Palmieri, C., Magi, G., Creti, R., Baldassarri, L., Imperi, M., Gherardi, G. & Facinelli, B. Interspecies mobilization of an erm(T)-carrying plasmid of *Streptococcus dysgalactiae* subsp. equisimilis by a coresident ICE of the ICESa2603 family. *Journal of Antimicrobial Chemotherapy* **68,** 23-26 (2013).
32. Seo, H.S., Xiong, Y.Q., Mitchell, J., Seepersaud, R., Bayer, A.S. & Sullam, P.M. Bacteriophage lysin mediates the binding of *Streptococcus mitis* to human platelets through interaction with fibrinogen. *PLoS pathogens* **6,** e1001047 (2010).
33. Hsieh, Y.C., Lin, T.L., Lin, C.M. & Wang, J.T. Identification of PblB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone. *Sci Rep* **5,** 12265 (2015).
34. Maruvada, R. & Kim, K.S. Extracellular Loops of the Eschericia coli Outer Membrane Protein A Contribute to the Pathogenesis of Meningitis. *The Journal of Infectious Diseases* **203,** 131-140 (2011).
35. Beverley, S.M., Owens, K.L., Showalter, M., Griffith, C.L., Doering, T.L., Jones, V.C. & McNeil, M.R. Eukaryotic UDP-Galactopyranose Mutase (GLF Gene) in Microbial and Metazoal Pathogens. *Eukaryotic Cell* **4,** 1147-1154 (2005).
36. Sitkiewicz, I., Stockbauer, K.E. & Musser, J.M. Secreted bacterial phospholipase A2 enzymes: better living through phospholipolysis. *Trends Microbiol* **15,** 63-9 (2007).
37. Nagiec, M.J., Lei, B., Parker, S.K., Vasil, M.L., Matsumoto, M., Ireland, R.M., Beres, S.B., Hoe, N.P. & Musser, J.M. Analysis of a Novel Prophage-encoded Group A Streptococcus Extracellular Phospholipase A2. *Journal of Biological Chemistry* **279,** 45909-45918 (2004).
38. Radin, J.N., Orihuela, C.J., Murti, G., Guglielmo, C., Murray, P.J. & Tuomanen, E.I. β-Arrestin 1 Participates in Platelet-Activating Factor Receptor-Mediated Endocytosis of *Streptococcus pneumoniae*. *Infection and Immunity* **73,** 7827-7835 (2005).
39. Cundell, D.R., Gerard, N.P., Gerard, C., Idanpaan-Heikkila, I. & Tuomanen, E.I. *Streptococcus pneumoniae* anchor to activated human cells by the receptor for platelet-activating factor. *Nature* **377,** 435-8 (1995).
40. Shukla, S.D. Platelet-activating factor receptor and signal transduction mechanisms. *The FASEB Journal* **6,** 2296-301 (1992).
41. Iovino, F., Seinen, J., Henriques-Normark, B. & van Dijl, J.M. How Does *Streptococcus pneumoniae* Invade the Brain? *Trends Microbiol* (2016).
42. Banks, D.J., Lei, B. & Musser, J.M. Prophage Induction and Expression of Prophage-Encoded Virulence Factors in Group A Streptococcus Serotype M3 Strain MGAS315. *Infection and Immunity* **71,** 7079-7086 (2003).
43. Kittang, B.R., Bruun, T., Langeland, N., Mylvaganam, H., Glambek, M. & Skrede, S. Invasive group A, C and G streptococcal disease in western Norway: virulence gene profiles, clinical features and outcomes. *Clinical Microbiology and Infection* **17,** 358-364 (2011).

44. Arrecubieta, C., Lopez, R. & Garcia, E. Molecular characterization of cap3A, a gene from the operon required for the synthesis of the capsule of *Streptococcus pneumoniae* type 3: sequencing of mutations responsible for the unencapsulated phenotype and localization of the capsular cluster on the pneumococcal chromosome. *J Bacteriol* **176**, 6375-83 (1994).
45. Cartee, R.T., Forsee, W.T., Jensen, J.W. & Yother, J. Expression of the *Streptococcus pneumoniae* type 3 synthase in Escherichia coli. Assembly of type 3 polysaccharide on a lipid primer. *J Biol Chem* **276**, 48831-9 (2001).
46. Waite, R.D., Penfold, D.W., Struthers, J.K. & Dowson, C.G. Spontaneous sequence duplications within capsule genes cap8E and tts control phase variation in *Streptococcus pneumoniae* serotypes 8 and 37. *Microbiology* **149**, 497-504 (2003).
47. Aanensen, D.M., Mavroidi, A., Bentley, S.D., Reeves, P.R. & Spratt, B.G. Predicted Functions and Linkage Specificities of the Products of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci. *Journal of Bacteriology* **189**, 7856-7876 (2007).
48. Brueggemann, A.B., Griffiths, D.T., Meats, E., Peto, T., Crook, D.W. & Spratt, B.G. Clonal Relationships between Invasive and Carriage *Streptococcus pneumoniae* and Serotype- and Clone-Specific Differences in Invasive Disease Potential. *Journal of Infectious Diseases* **187**, 1424-1432 (2003).
49. Saha, S.K., Naheed, A., El Arifeen, S., Islam, M., Al-Emran, H., Amin, R., Fatima, K., Brooks, W.A., Breiman, R.F., Sack, D.A., Luby, S.P. & Group, P.S. Surveillance for Invasive *Streptococcus pneumoniae* Disease among Hospitalized Children in Bangladesh: Antimicrobial Susceptibility and Serotype Distribution. *Clinical Infectious Diseases* **48**, S75-S81 (2009).
50. Shakoor, S., Kabir, F., Khowaja, A.R., Qureshi, S.M., Jehan, F., Qamar, F., Whitney, C.G. & Zaidi, A.K.M. Pneumococcal Serotypes and Serogroups Causing Invasive Disease in Pakistan, 2005–2013. *PLoS ONE* **9**, e98796 (2014).
51. Attarpour - yazdi, M.M., Ghamarian, A., Mousaviehzadeh, M. & Davoudi, N. Identification of the serotypes of bacterial meningitis agents; implication for vaccine usage. *Iranian Journal of Microbiology* **6**, 211-218 (2014).
52. Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P. & Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45**, 656-63 (2013).
53. Iroh Tam, P.-Y. & Young, M.E. Serotype 35B Pneumococcal Meningitis in an Infant: Effect of Conjugate Vaccines on Invasive Disease and Implications for Practice. *Clinical Pediatrics* (2015).
54. Ardanuy, C., Marimon, J.M., Calatayud, L., Gimenez, M., Alonso, M., Grau, I., Pallares, R., Perez-Trallero, E. & Linares, J. Epidemiology of invasive pneumococcal disease in older people in Spain (2007-2009): implications for future vaccination strategies. *PLoS One* **7**, e43619 (2012).
55. Camilli, R., Daprai, L., Cavrini, F., Lombardo, D., D'Ambrosio, F., Del Grosso, M., Vescio, M.F., Landini, M.P., Pascucci, M.G., Torresani, E., Garlaschi, M.L., Sambri, V. & Pantosti, A. Pneumococcal Carriage in Young Children One Year after Introduction of the 13-Valent Conjugate Vaccine in Italy. *PLoS ONE* **8**, e76309 (2013).
56. Domenech, M., Damián, D., Ardanuy, C., Liñares, J., Fenoll, A. & García, E. Emerging, Non-PCV13 Serotypes 11A and 35B of *Streptococcus pneumoniae* Show High Potential for Biofilm Formation In Vitro. *PLoS ONE* **10**, e0125636 (2015).
57. Kaplan, S.L., Barson, W.J., Lin, P.L., Romero, J.R., Bradley, J.S., Tan, T.Q., Hoffman, J.A., Givner, L.B. & Mason, E.O.J. Early Trends for Invasive Pneumococcal Infections in

Children After the Introduction of the 13-valent Pneumococcal Conjugate Vaccine. *The Pediatric Infectious Disease Journal* **32**, 203-207 (2013).
58. Lopez, R. & Garcia, E. Recent trends on the molecular biology of pneumococcal capsules, lytic enzymes, and bacteriophage. *FEMS Microbiol Rev* **28**, 553-80 (2004).
59. Flores, C.O., Meyer, J.R., Valverde, S., Farr, L. & Weitz, J.S. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* **108**, E288-E297 (2011).
60. DeBardeleben, H.K., Lysenko, E.S., Dalia, A.B. & Weiser, J.N. Tolerance of a Phage Element by *Streptococcus pneumoniae* Leads to a Fitness Defect during Colonization. *Journal of Bacteriology* **196**, 2670-2680 (2014).
61. Croucher, N.J., Walker, D., Romero, P., Lennard, N., Paterson, G.K., Bason, N.C., Mitchell, A.M., Quail, M.A., Andrew, P.W., Parkhill, J., Bentley, S.D. & Mitchell, T.J. Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone *Streptococcus pneumoniae* Spain23F ST81. *Journal of Bacteriology* **191**, 1480-1489 (2009).
62. Zgur-Bertok, D. DNA damage repair and bacterial pathogens. *PLoS Pathog* **9**, e1003711 (2013).
63. Rubins, J.B. & Janoff, E.N. Pneumolysin: A multifunctional pneumococcal virulence factor. *Journal of Laboratory and Clinical Medicine* **131**, 21-27 (1998).
64. Cockeran, R., Anderson, R. & Feldman, C. The role of pneumolysin in the pathogenesis of *Streptococcus pneumoniae* infection. *Current Opinion in Infectious Diseases* **15**, 235-239 (2002).
65. Bensing, B.A., Rubens, C.E. & Sullam, P.M. Genetic Loci of *Streptococcus mitis* That Mediate Binding to Human Platelets. *Infection and Immunity* **69**, 1373-1380 (2001).
66. Bensing, B.A., Siboo, I.R. & Sullam, P.M. Proteins PblA and PblB of *Streptococcus mitis*, Which Promote Binding to Human Platelets, Are Encoded within a Lysogenic Bacteriophage. *Infection and Immunity* **69**, 6186-6192 (2001).
67. Gladstone, R.A., Jefferies, J.M., Tocheva, A.S., Beard, K.R., Garley, D., Chong, W.W., Bentley, S.D., Faust, S.N. & Clarke, S.C. Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine* **33**, 2015-21 (2015).
68. Elberse, K.E., van der Heide, H.G., Witteveen, S., van de Pol, I., Schot, C.S., van der Ende, A., Berbers, G.A. & Schouls, L.M. Changes in the composition of the pneumococcal population and in IPD incidence in The Netherlands after the implementation of the 7-valent pneumococcal conjugate vaccine. *Vaccine* **30**, 7644-51 (2012).
69. Cremers, A.J., Kokmeijer, I., Groh, L., de Jonge, M.I. & Ferwerda, G. The role of ZmpC in the clinical manifestation of invasive pneumococcal disease. *Int J Med Microbiol* **304**, 984-9 (2014).
70. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575-1584 (2002).
71. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).
72. Wernersson, R. & Pedersen, A.G. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* **31**, 3537-9 (2003).
73. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
74. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).

Chapter 7

Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease

Rahajeng N. Tunjungputri*
Fredrick M. Mobegi*
Amelieke J. Cremers
Christa E. van der Gaast – de Jongh
Gerben Ferwerda
Jacques F. Meis
Nel Roeleveld
Stephen D. Bentley
Sacha A.F.T. van Hijum
Andre J. van der Ven
Quirijn de Mast
Aldert Zomer
Marien I. de Jonge

*authors contributed equally

Manuscript submitted for publication

Abstract

To improve our understanding about the severity of invasive pneumococcal disease (IPD), in this study we investigated the association between the genotype of *Streptococcus pneumoniae* and disease outcome of 349 bacteremic patients. Pneumococcal genome-wide association (GWAS) analysis demonstrated a strong correlation between 30-day mortality and the presence of the phage-derived *pblB* gene, encoding a platelet-binding protein of which consequences on platelet activation was previously unknown. Platelets are increasingly recognized as key players of the innate immune system, and in sepsis, excessive platelet activation contributes to microvascular obstruction, tissue hypoperfusion and finally multi-organ failure leading to mortality. Our *in vitro* studies revealed that *pblB* expression was induced by fluoroquinolones but not by the beta-lactam antibiotic Penicillin G. Subsequently, we determined *pblB* induction and platelet activation by incubating whole blood with wild type or knock-out mutant of *pblB*, in the presence or absence of antibiotics commonly administered in our patient cohort. *PblB*-dependent enhancement of platelet activation, as measured by increased expression of the alpha-granule protein P-selectin, the binding of fibrinogen to the activated αIIbβ3 receptor and the formation of platelet-monocyte complex, occurred irrespective of the antibiotics exposure. In conclusion, the presence of *pblB* on the pneumococcal chromosome potentially leads to increased mortality in patients with an invasive *S. pneumoniae* infection, which may be explained by enhanced platelet activation. This study highlights the clinical utility of bacterial GWAS, followed by functional characterization to identify bacterial factors involved in disease severity.

Importance

The exact mechanisms causing mortality in invasive pneumococcal disease (IPD) patients is not completely understood. We examined 349 patients with IPD and found in a bacterial genome-wide association study (GWAS) that the presence of the phage-derived gene *pblB* was associated with mortality in the first 30 days after hospitalization. Although *pblB* had been extensively studied in *Streptococcus mitis*, its consequence on the interaction between platelets and *Streptococcus pneumoniae* is largely unknown. Platelets are important in immunity and inflammation, and excessive platelet activation contributes to microvascular obstruction and multi-organ failure, leading to mortality. We therefore developed this study to assess whether the phage-derived pblB might increase the risk for mortality in IPD patients through its effect on enhanced platelet activation. This study would also show the value of integrating extensive bacterial genomics and clinical data in predicting and understanding pathogen virulence, which in turn will help to improve prognosis and therapy.

Introduction

Streptococcus pneumoniae or the pneumococcus is a frequent colonizer of the nasopharynx. In a minority of carriers, infection progresses to pneumococcal disease with an estimated 1.6 million deaths annually [1, 2]. The largest clinical burden of invasive pneumococcal disease (IPD) is seen in young children and older adults, who present mostly with sepsis and meningitis. Case mortality rates are estimated to range from 11 to 30% in adults [3-5], with treatment becoming complicated due to the worldwide emergence of multi-drug resistance [6]. Therefore, it is of utmost importance to fully understand the pathogenic mechanisms of pneumococcal disease in order to improve treatment and prognosis of critically ill patients.

Recently, the utilization of whole genome sequencing and analyses for predicting and understanding pathogen virulence has been highlighted [7]. In this study, we performed a genome-wide association analysis (GWAS) on 349 pneumococcal draft genomes of blood isolates from patients who were admitted with IPD in two Dutch hospitals. We identified a significant association between 30-day mortality and the presence of *pblB*, encoding for a platelet binding protein that was also reported to function in adhesion, which might explain clonal success [8]. In a subsequent functional study, we investigated the induction of the phage-derived *pblB* expression by fluoroquinolones in *S. pneumoniae*. Lastly, we simulated the *in vivo* conditions using an *ex vivo* whole blood assay demonstrating the importance of *pblB* in enhancing platelet activation.

Platelets are an important part of the innate immune system, and can interact with and be activated by *S. pneumoniae*. In sepsis, platelet activation and platelet-leukocyte complex formation contribute to microvascular obstruction, tissue hypoperfusion and finally multi-organ failure [9]. The role of this phage-derived gene in the clinical outcome and severity of IPD patients and its consequences for platelet activation warrant further study.

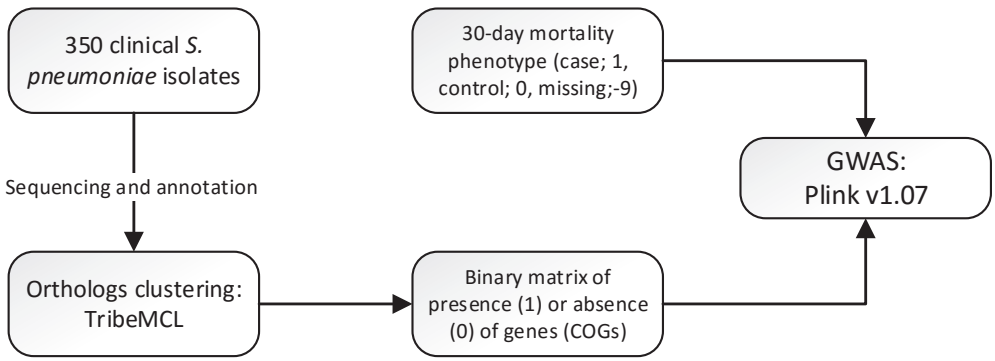


Figure 1. Flowchart of computational method used to identify the association of presence of the *pblB* gene with 30-day mortality

Results

PblB is an independent determinant of 30-day mortality in IPD patients

We conducted an unbiased association study for the presence or absence of genes and mortality within the first 30 days of hospitalization (Figure 1). Analysis was performed using 349 pneumococcal strains collected from a clinical IPD cohort in Nijmegen which comprised of strains from multiple lineages [10]. The overall 30-day mortality within this IPD cohort was 11% (37/346, outcome unknown for 3 cases). We observed that out of the 1946 orthologous genes (OGs) of the pneumococcal accessory genome, the presence of the phage-encoded gene *pblB* (OG_17), was positively associated with 30-day mortality in IPD patients (GWAS stratified for population structure with BAPS, p=0.00034 Bonferroni corrected for multiple testing) [11]. All the OGs associated with 30-day mortality with corrected *p*-values <0.05 are listed in Table S1. The *pblB* gene was present in 48% of the 349 clinical strains and co-occurred with other genes predicted to encode for phage components or hypothetical proteins. However, we identified *pblB* as the phage-derived gene having the strongest statistical correlation with 30-day mortality.

Among the IPD cases caused by pneumococci containing the *pblB* gene (*pblB*$^+$), 27 out of 165 died within 30 days (16.4%), compared to only 10 out of 181 (5.5%) caused by pneumococci not containing the *pblB* gene (*pblB*$^-$) (p=0.0011; OR 3.3). In a sub-analysis of cases who died without any limitations of medical treatment, 30-day mortality was 15/165 (9.1%) in *pblB*$^+$ and 6/171 (3.3%) in *pblB*$^-$ cases, which remained statistically significant (p=0.022; OR 2.8). For all cases, the presence of *pblB* was an independent determinant of 30-day mortality (OR 3.4, 95% CI: 1.5-7.6), next to Charlson comorbidity index score (OR 1.5, 95% CI: 1.2-1.7) and meningitis (OR 4.6, 95% CI: 1.6-13.7). For pneumonia cases separately, in addition to PSI score (OR 1.4, 95% CI: 1.1-1.7) and Charlson comorbidity score (OR 1.02, 95% CI: 1.01-1.04), both designed to predict mortality, the presence of *pblB* was an independent risk factor for 30-day mortality (OR 3.4, 95% CI: 1.2-9.5).

Fluoroquinolones induced the expression of *pblB*

It was unknown whether *pblB*-containing temperate pneumophages are specifically induced by this group of antibiotics *in vitro*. Therefore, different doses of ciprofloxacin (CPX) and levofloxacin (LVX; both belonging to the fluoroquinolone group of antibiotics), mitomycin C (MitC) and penicillin G (PenG; a beta-lactam antibiotic) were tested on three *pblB*-containing pneumococcal strains (PBCN0103, PBCN0226, PBCN0239) in THY medium to determine the sublethal dose of the four antibiotics (data not shown). To confirm that the selected doses were not bactericidal, the number colony forming units (CFU) were determined after exposure of MitC, PenG and the fluoroquinolones for two hours at 37 °C and 5% CO_2 (Figure 2A). At the same time point, the difference in expression of *pblB*, relative to *gyrA,* was measured. The DNA cross-linking agent MitC was included as positive control. Both the fluoroquinolones CPX and LVX induced the

expression of *pblB*, the latter being the strongest inducer, which appeared specific for this group of antibiotics, as the beta-lactam antibiotic PenG did not induce the expression. Furthermore, strong variation was found between the different pneumococcal strains (Figure 2B).

Simulation of the clinical conditions in a whole blood *ex vivo* assay

Of the 312 patients with sequenced strains and known empirical treatment, 28% (n=88) received only beta-lactam, 4% (n=11) received only fluoroquinolones, and 44% received a combination of a beta-lactam and a fluoroquinolone. To simulate the aforementioned clinical conditions, we incubated live pneumococci strain PBCN0162 containing a mutationally inactivated *pblB* gene (Δ*pblB*) or wild type (wt), with and without antibiotics (PenG, CPX, and a combination of PenG and CFX) in whole blood, determined the expression of *pblB* using qPCR (Figure 3A) and measured in the same samples the activation of platelets. We were able to measure *pblB* expression of the wt pneumococci in the whole blood samples without antibiotics and its increase in the presence of antibiotics (mean CQ value 30.6, 95% confidence interval 29.5-31.7). We first analyzed whether the different antibiotics significantly affect the wt-/Δ*pblB*-bacteria-mediated platelet activation state in whole blood using a liner mixed model. We found that in all cases, stronger activation of platelets was observed with wt pneumococci as compared to Δ*pblB*, which clearly indicates that PblB induces enhanced platelet activation irrespective of the exposure to antibiotics (Figures 3B).

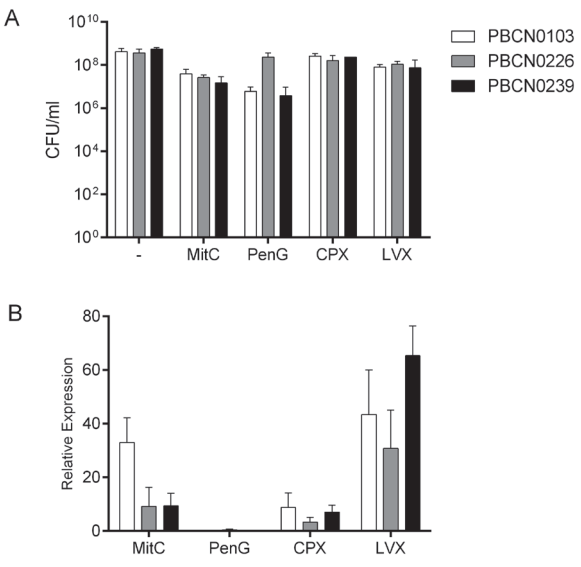


Figure 2. Sublethal doses of antibiotics induced pneumococcal expression of the *pblB* phage in culture medium. Average CFU values were determined after incubation of the three clinical pneumococcal strains (PBCN0103, PBCN0226, PBCN0239) for two hours at 37°C and 5% CO_2 in THY medium with sub-lethal doses of antibiotics; mitomycin C (MitC), penicillin G (PenG), ciprofloxacin (CPX) and levofloxacin (LVX). The condition without antibiotics (-) was included as negative control (A). Induction of *pblB* expression after two hours of incubation with sub-lethal doses of antibiotics were determined by qRT-PCR measuring levels of mRNA relative to the control *gyrA* (B). The data shown represent the means with standard deviations of three independently performed experiments.

Whilst PenG did not strongly induce expression of *pblB* in THY-medium-grown pneumococci (Figure 2B), we observed PenG-dependent induction (~3-fold) of expression in whole blood (Figure 3A). This might be caused by an indirect effect as a consequence of the bactericidal effect of PenG leading to the production of reactive oxygen species (ROS) which has DNA damaging effects, inducing the expression of *pblB*. Despite the fact that expression of *pblB* was much stronger in whole blood containing CPX, platelet activation is not increased accordingly, indicating a close to maximum activation under these conditions

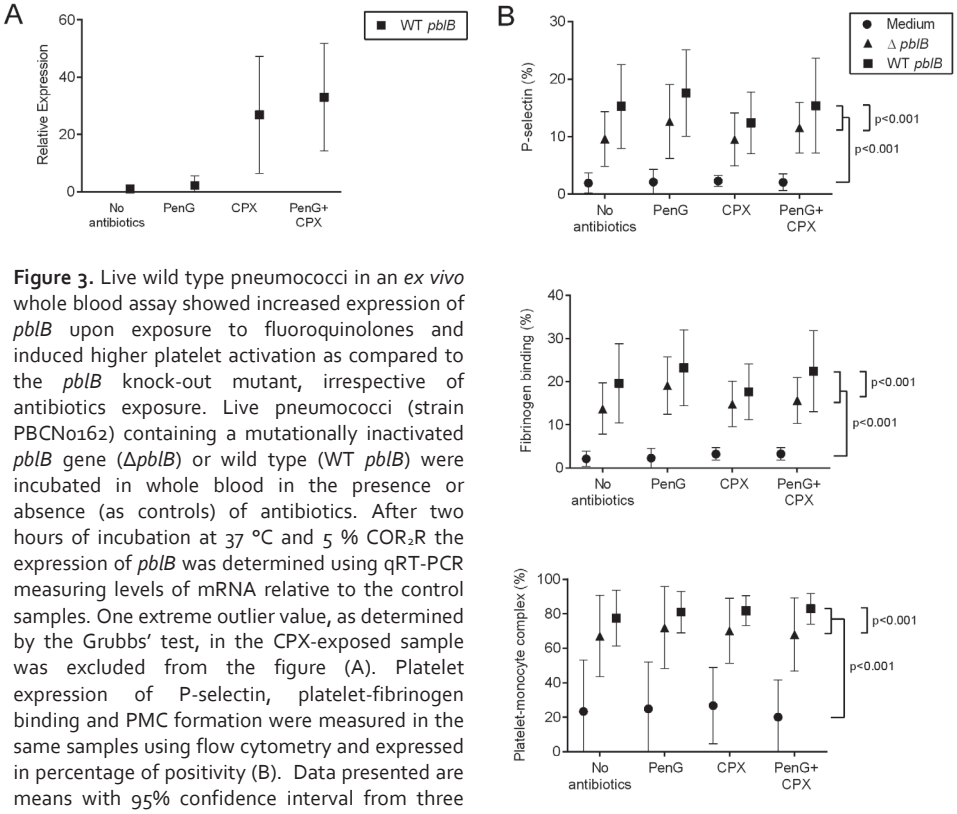


Figure 3. Live wild type pneumococci in an *ex vivo* whole blood assay showed increased expression of *pblB* upon exposure to fluoroquinolones and induced higher platelet activation as compared to the *pblB* knock-out mutant, irrespective of antibiotics exposure. Live pneumococci (strain PBCN0162) containing a mutationally inactivated *pblB* gene (Δ*pblB*) or wild type (WT *pblB*) were incubated in whole blood in the presence or absence (as controls) of antibiotics. After two hours of incubation at 37 °C and 5 % COR_2R the expression of *pblB* was determined using qRT-PCR measuring levels of mRNA relative to the control samples. One extreme outlier value, as determined by the Grubbs' test, in the CPX-exposed sample was excluded from the figure (A). Platelet expression of P-selectin, platelet-fibrinogen binding and PMC formation were measured in the same samples using flow cytometry and expressed in percentage of positivity (B). Data presented are means with 95% confidence interval from three independent experiments with blood derived from a total of 6 human volunteers. Differences were analyzed using a linear mixed model, with the Bonferroni correction for multiple testing. p<0.05 was considered statistically significant. PenG, Penicillin G; CPX, ciprofloxacin.

Discussion

In the present study, GWAS was performed, in an unbiased manner, using the sequences of 349 *S. pneumoniae* invasive disease isolates to test for associations between the presence or absence of genes in the pneumococcal accessory genome and 30-day mortality.

The presence of the phage-encoded *pblB* gene was positively associated with 30-day mortality in patients with IPD. This finding suggested the role of *pblB* in the pathogenesis and expected cause of death of IPD. The presence of the *pblB* phage gene as risk factor remained after adjustment for the local pneumococcal population structure. We therefore speculate that similar studies in other areas with different pneumococcal populations would yield similar findings, although this requires confirmation by other studies. Interestingly, the *pblB* phage in our cohort was barely present in serotypes 1 and 7F, which are associated with a lower risk of death than other serotypes [12].

Up to 75% of pneumococcal clinical isolates have been shown to carry bacteriophages (pneumophages) [13], which may be distributed among pneumococcal isolates with different capsular serotypes, indicating that these mobile genetic elements are widely spread among clinically relevant pneumococcal strains [14]. The hypothesis that bacteria acquire virulence properties from phages is widely accepted [15], however, there has been a paucity of data supporting the role of bacteriophages in the pathogenesis of *S. pneumoniae*-caused diseases.

The PblB protein was first characterized in *S. mitis* as a surface protein encoded by a lysogenic bacteriophage (SM1) and described to mediate the binding to human platelets [16]. In *S. mitis*, induction of the lytic cycle leads to permeabilization of the cell wall with subsequent release and expression of PblB on the cell surface, allowing the interaction with and propagation of platelet activation. The expression of PblB in *S. mitis* was also found to contribute to virulence in an *in vivo* rabbit model of infective endocarditis [17]. These findings demonstrate that *pblB* might have an important role in endovascular infection and recently, PblB was reported to mediate pneumococcal adhesion to platelets and lung epithelial cells [8]. However, the functional consequences of *pblB* expression in *S. pneumoniae* and its role in the pathogenesis of pneumococcal disease is largely unknown.

Most patients in this cohort were treated with a combination of penicillin and ciprofloxacin, which represents a common first line empiric antibiotic regime for severe community acquired pneumonia in the Netherlands [18]. We therefore proceeded with *ex vivo* experiments in which live pneumococci were incubated in whole blood supplemented with penicillin or ciprofloxacin or a combination of the two, to simulate the clinical conditions. The wild type pneumococci clearly demonstrated enhanced platelet activation. Interestingly, there were differences in platelet activation between knock-out mutant and wild type pneumococci even in the absence of high *pblB* induction by the

antibiotics. This may be explained by a constitutive expression of *pblB*, which despite its low level was sufficient to induce platelet activation, as had been described in *S. mitis* [17].

S. pneumoniae has been shown to directly activate platelets mainly through TLR2 [19], with FcγRIIA and integrin αIIbβ3 being involved in the amplification of bacteria-induced platelet activation [20]. This leads to platelet degranulation and subsequently the release of an array of chemokines and inflammatory mediators which may modulate not only their own function but also cells around them [21, 22]. Our findings that whole blood exposure to WT pneumococci result in higher platelet activation compared to the *pblB* knock-out mutant may explain why bacteremic patients, infected with pneumococci containing the *pblB* gene, have a higher chance to die within 30 days. An approximate of 20% increase from baseline values in platelet P-selectin expression and PMC has been associated with adverse cardiovascular events and the acute phase of ischaemic stroke [23, 24]. The increase in platelet activation associated with *pblB* in our *ex vivo* assays exceeded this aforementioned value. By causing enhanced platelet activation, bacteria can become engulfed in a septic thrombus and protected from other cells of the immune system, allowing them to persist in the circulation [25]. We speculate that the *pblB*-enhanced platelet activation may confer this survival advantage for *S. pneumoniae*. On the other hand, the resulting excess of platelet activation together with platelet clumping, platelet-leukocyte and platelet-endothelium aggregation and increased fibrin formation result in enhanced thrombo-inflammatory responses, microvascular obstruction, tissue hypoperfusion and finally multi-organ failure in sepsis [26, 27]. The increase of PMC formation predicts mortality in older septic patients [28], and platelet consumption associated with platelet activation in sepsis patients leads to thrombocytopenia, which has been shown to increase the risk of mortality [29-31].

Our results have several potential clinical implications. Firstly, that *pblB* is associated with 30-day mortality suggests the potential of bacterial GWAS for improvement of clinical management of IPD patients. Secondly, our results demonstrated that fluoroquinolones induce higher *pblB* expression. However, the presence of fluoroquinolones was not required by the *pblB*-expressing wild type pneumococci to enhance platelet activation when compared with the knock-out mutant. Given that fluoroquinolones are frequently used in the management of community-acquired pneumonia for the coverage of atypical pathogens [32], sufficiently-powered studies are needed to investigate the clinical outcomes of the interplay between antibiotics regimen and *pblB* before drawing any conclusions. Thirdly, our study further highlights the importance of platelet-bacterial interaction and platelet activation, both in providing a survival advantage for bacteria and in posing increased risk of mortality in patients. There is more and more data on the use of platelet function inhibitors in sepsis, however, these results at times contradict [9]. Platelet inhibition by the $P2Y_{12}$ receptor antagonists reduces the release of pro-inflammatory mediators from the platelet α-granules [33]. Taken together with our findings, the role of anti-platelet agents as adjunctive therapy in sepsis warrants further investigation. To the best of our knowledge, this is the only patient-based study which

independently and in an unbiased manner reveal the role of *pblB* in the pathogenesis of IPD based on a vast analysis of both genomics and clinical data, adding substantial evidence to only two recent studies on pneumococcal *pblB in vitro* and in mice [8, 34].

In conclusion, we have integrated genome sequencing and GWAS with functional characterization to investigate the clinical role of *pblB* in the mortality of patients with IPD. Bacterial GWAS may be an important tool to study the potential predictive value of certain virulence genes. As genomic sequencing is increasingly being utilized, we believe that this integrated approach will assist greatly in elucidating the mechanisms of bacterial pathogenesis leading to the development of novel diagnostics and new therapeutic approaches.

Materials and Methods

Study population

Consecutive patients hospitalized with a bacteremic pneumococcal infection at two Dutch hospitals between 2001 and 2011 were included in the study. Detailed clinical data were obtained on patient characteristics, clinical severity, treatment and the course of disease. Corresponding blood culture isolates of *S. pneumoniae* were collected and serotyped as described before [10]. For 349 of the isolated strains, sequencing, assembly of draft genomes and annotation was determined as previously described [35]. This study was reviewed and approved by the Local Medical Ethical Committees. All adult patients and healthy volunteers involved in this study provided written informed consent.

Orthologous clustering and GWAS

Orthologous genes (OGs) from *S. pneumoniae* used in this study have previously been described by our group [35]. Putative protein coding sequences were investigated using an "all-versus-all" protein BLAST (blastP), with a 10e-15 *e*-value cut-off and a BLOSUM90 substitution matrix. The results were subsequently clustered into clusters of orthologous groups using TribeMCL [35, 36], resulting into a total of 3021 orthologous genes (OGs), 1075 of which were conserved in all isolates in a single copy. The population (sub)-structure (sequence clusters; SCs) used for population stratification in the study have also been previously characterized [35]. Basing disease severity on mortality within the first 30 days of admission to the hospital, the pneumococcal isolates were categorized into three categories: derived from patients who died (n=37), from patients who survived (n=309), and from patients of whom the data was not captured (n=3). The Cochran-Mantel-Haenszel (CMH) association statistics was employed to test the associations between the presence or absence of pneumococcal OGs and 30-day mortality, conditional on the bacterial population substructure as proposed by BAPS [11]. All associations were

determined using PLINK [37]. Candidate OGs were selected based on association test with p <0.05 (Bonferroni adjusted for multiple testing).

Adjustment for covariates of mortality

Certain patients had predetermined limitations of medical treatments, for example, opted not to be transferred to the intensive care unit. Therefore, the relation between OGs and 30-day mortality was also established separately for those who died after fully-applied treatment. Potentially interesting covariates of 30-day mortality, in addition to the pneumococcal OG, were selected by their distribution in relation to 30-day mortality. The variables considered were gender, age, year of inclusion, comorbidities (i.e. cancer, COPD, diabetes mellitus, liver-, renal-, cardiovascular- and cerebrovascular disease, Charlson comorbidity index score), clinical diagnosis, blood C-reactive protein level, presence of Systemic Inflammatory Response Syndrome (SIRS) and pleural effusion, Pneumonia Severity Index (PSI) score, and class of antibiotics administered. The final model for 30-day mortality as a dependent variable was constructed via binary logistic regression analysis by likelihood ratio based backward modeling, entering the pneumococcal OG plus identified possible covariates as explaining variables. Analyses of covariates were performed using IBM SPSS statistics 23.

Induction of *pblB* expression by antibiotics

Three isolates derived from the group of deceased patients, containing the *pblB* gene, were selected: PBCN0103, PBCN0226 and PBCN0239. Different concentrations of mitomycin C, penicillin G, ciprofloxacin and levofloxacin (all purchased from Sigma-Aldrich, Zwijndrecht, The Netherlands) were tested to determine the sub-lethal doses. The pneumococci were grown in THY medium to midlog (OD 0.3), then diluted to OD 0.1, supplemented with 0.132 µg/ml mitomycin C, 0.0125 µg/ml penicillin G, 0.533 µg/ml ciprofloxacin or 0.533 µg/ml levofloxacin, and grown for an additional two hours at 37°C with 5% CO_2. Subsequently, serial dilutions were incubated on blood agar plates (BD) and incubated overnight at 37°C with 5% CO_2. Experiments were performed in triplicate to determine the expression of *pblB*. Mitomycin C was included as positive control, as it was previously shown to induce *pblB* expression [38]. Penicillin G was included as negative control. After two hours of growth, pneumococci were harvested by centrifugation. The supernatant was discarded and a 2:1 volume of RNA protect (Qiagen, Hilden, Germany) was added to the pellet. RNA was isolated using the RNeasy kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. Residual DNA was removed with a DNAse treatment using the Ambion Turbo DNA-free kit according to manufacturer's instructions (Ambion, Austin, TX, USA). The qRT-PCR was performed as previously described by DeBardeleben *et al.* [38] using the following primers: HBgyrAF: AATGAACGGGAACCCTTGGT, HBgyrAR: CCATCCCAACCGCGATAC, pblB_F: TACAGCTGTGAAAGCCTTGG, pblB_R: GATAGCCATCTGGATTCTCAGG.

Construction of *S. pneumoniae* strain PBCN0162Δ*pblB*

A directed gene deletion mutant of *S. pneumoniae* strain PBCN0162 was generated by allelic exchange of the target gene (*pblB*) with a spectinomycin resistance cassette (obtained from pR412T7), using the megaprimer polymerase chain reaction method, this resulted in PBCN0162Δ*pblB*. Briefly, flanking regions of ~500 bp, containing less than 150 bp of the coding sequence of the target genes, were amplified by PCR, with chromosomal DNA as the template. For each flanking region, the primer closest to the target gene (extension plus _L2 or _R2) contained an additional sequence complementary to primer PBpR412_L or PBpR412_R. In a second PCR, the PCR products of the two flanking regions and the antibiotic resistance cassette were combined, leading to incorporation of the antibiotic resistance cassette between the two flanking regions of the target gene, as previously described by Burghout et al., 2007 [39]. The primer sequences are provided in Table S1 in the supplemental material. Subsequently, the megaprimer PCR product was used for transformation of competent PBCN0162. Mutants, selected on blood agar plates containing spectinomycin, were assessed by colony PCR for recombination at the desired location on the chromosome. Chromosomal DNA was isolated from the mutants and used for transformation of competent strain PBCN0162. Gene inactivation was confirmed by quantitative real-time PCR gene expression analyses as described above (see '*Induction of pblB expression by antibiotics*').

***Ex vivo* (whole blood) assays**

Whole blood was obtained from healthy volunteers (n=6) after informed consent using 3.2% citrate-anticoagulated tubes (BD Vacutainer, Becton Dickinson, Plymouth, UK) and exposed to $1x10^7$ CFU/ml Δ*pblB* or wt pneumococci for 30 minutes at 37°C. Subsequently, either medium, penG (0.0125 ug/ml), CPX (0.533 ug/ml), or a combination of penG and CPX were added, and samples were incubated for 2 hrs at 37°C. RNA isolation and qRT-PCR was performed as described in the previous section. These whole blood samples were also collected for measurement of platelet activation and PMC by flow cytometry.

Measurement of platelet activation and platelet-monocyte (PMC) complex formation by flow cytometry

Platelet activation was measured by whole blood flow cytometry as previously described [40] by quantifying the platelet membrane expression of the α-granule protein P-selectin (CD62P) and the binding of fibrinogen to the activated αIIbβ3 receptor (GPIIbIIIa complex). The following antibodies were used to incubate samples from the whole blood *ex vivo* assay: PE-labelled anti-CD62P (Bio-Legend, San Diego, CA, USA), FITC-labelled anti-fibrinogen (F0111-FITC; DAKO Ltd., High Wycombe, UK) and PC7-labelled anti-CD61 (platelet glycoprotein IIIa, Beckman Coulter, Miami, FL, USA), the latter as platelet identification marker. The percentage of CD62P and fibrinogen on CD61-positive events were determined. Formation of PMC was measured by incubating samples with PC7-

labelled anti-CD61 and PE-labelled anti-CD14 (a-glycosylphosphatidylinositol (GPI)-linked membrane glycoprotein; Bio-Legend). After 20 min incubation, Optilyse B (Beckman Coulter, Fullerton CA, USA) was added to lyse erythrocytes. PMC formation was determined by quantifying the MFI of CD14+ cells that were also positive for the platelet identification marker CD61. All samples were measured using a FC500 flow cytometer (Beckman Coulter).

Statistical Analyses

Results from independent experiments (involving n = 6 donors) were pooled, data are provided as means with 95% confidence interval. For data from the *ex vivo* whole blood assay, we used a linear mixed model approach to analyze the effects of the interaction between bacterial exposure and the different antibiotics on platelet activation (SPSS, Chicago IL, USA). The level of significance was set at $p<0.05$.

Acknowledgements

We thank Dr. Ton de Han for the statistical support.

References

1. Rajaratnam, J.K., Marcus, J.R., Flaxman, A.D., Wang, H., Levin-Rector, A., Dwyer, L., Costa, M., Lopez, A.D. & Murray, C.J. Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards Millennium Development Goal 4. *The Lancet* **375,** 1988-2008 (2010).
2. World Health Organization. Pneumococcal disease. Vol. 2015 (2005).
3. Hung, I.F.-N., Tantawichien, T., Tsai, Y.H., Patil, S. & Zotomayor, R. Regional epidemiology of invasive pneumococcal disease in Asian adults: epidemiology, disease burden, serotype distribution, and antimicrobial resistance patterns and prevention. *International Journal of Infectious Diseases* **17,** e364-e373 (2013).
4. Mufson, M.A. & Stanek, R.J. Bacteremic pneumococcal pneumonia in one American city: a 20-year longitudinal study, 1978–1997. *The American Journal of Medicine* **107,** 34-43 (1999).
5. Rock, C., Sadlier, C., Fitzgerald, J., Kelleher, M., Dowling, C., Kelly, S. & Bergin, C. Epidemiology of invasive pneumococcal disease and vaccine provision in a tertiary referral center. *European Journal of Clinical Microbiology & Infectious Diseases* **32,** 1135-1141 (2013).
6. Ginsburg, A.S., Tinkham, L., Riley, K., Kay, N.A., Klugman, K.P. & Gill, C.J. Antibiotic non-susceptibility among *Streptococcus pneumoniae* and *Haemophilus influenzae* isolates identified in African cohorts: a meta-analysis of three decades of published studies. *International journal of antimicrobial agents* **42,** 482-491 (2013).
7. Priest, N.K., Rudkin, J.K., Feil, E.J., Van Den Elsen, J.M.H., Cheung, A., Peacock, S.J., Laabei, M., Lucks, D.A., Recker, M. & Massey, R.C. From genotype to phenotype: can systems biology be used to predict *Staphylococcus aureus* virulence? *Nature Reviews Microbiology* **10,** 791-797 (2012).
8. Hsieh, Y.-C., Lin, T.-L., Lin, C.-M. & Wang, J.-T. Identification of PblB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone. *Scientific reports* **5**(2015).
9. de Stoppelaar, S.F., van 't Veer, C. & van der Poll, T. The role of platelets in sepsis. *Thromb Haemost* **112,** 666-77 (2014).
10. Cremers, A.J., Meis, J.F., Walraven, G., Jongh, C.E., Ferwerda, G. & Hermans, P.W. Effects of 7-valent pneumococcal conjugate 1 vaccine on the severity of adult 2 bacteremic pneumococcal pneumonia. *Vaccine* **32,** 3989-94 (2014).
11. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLoS Comput Biol* **5,** e1000455 (2009).
12. Weinberger, D.M., Harboe, Z.B., Sanders, E.A., Ndiritu, M., Klugman, K.P., Ruckinger, S., Dagan, R., Adegbola, R., Cutts, F., Johnson, H.L., O'Brien, K.L., Scott, J.A. & Lipsitch, M. Association of serotype with risk of death due to pneumococcal pneumonia: a meta-analysis. *Clin Infect Dis* **51,** 692-9 (2010).
13. Ramirez, M., Severina, E. & Tomasz, A. A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J Bacteriol* **181,** 3618-25 (1999).
14. Gindreau, E., Lopez, R. & Garcia, P. MM1, a temperate bacteriophage of the type 23F Spanish/USA multiresistant epidemic clone of *Streptococcus pneumoniae*: structural analysis of the site-specific integration system. *J Virol* **74,** 7803-13 (2000).

15. Flores, C.O., Meyer, J.R., Valverde, S., Farr, L. & Weitz, J.S. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* **108**, E288-E297 (2011).
16. Bensing, B.A., Siboo, I.R. & Sullam, P.M. Proteins PblA and PblB of *Streptococcus mitis*, Which Promote Binding to Human Platelets, Are Encoded within a Lysogenic Bacteriophage. *Infection and Immunity* **69**, 6186-6192 (2001).
17. Mitchell, J., Siboo, I.R., Takamatsu, D., Chambers, H.F. & Sullam, P.M. Mechanism of cell surface expression of the *Streptococcus mitis* platelet binding proteins PblA and PblB. *Molecular microbiology* **64**, 844-857 (2007).
18. Wiersinga, W.J., Bonten, M.J., Boersma, W.G., Jonkers, R.E., Aleva, R.M., Kullberg, B.J., Schouten, J.A., Degener, J.E., Janknegt, R., Verheij, T.J., Sachs, A.P. & Prins, J.M. SWAB/NVALT (Dutch Working Party on Antibiotic Policy and Dutch Association of Chest Physicians) guidelines on the management of community-acquired pneumonia in adults. *Neth J Med* **70**, 90-101 (2012).
19. Keane, C., Tilley, D., Cunningham, A., Smolenski, A., Kadioglu, A., Cox, D., Jenkinson, H.F. & Kerrigan, S.W. Invasive *Streptococcus pneumoniae* trigger platelet activation via Toll-like receptor 2. *J Thromb Haemost* **8**, 2757-65 (2010).
20. Arman, M., Krauel, K., Tilley, D.O., Weber, C., Cox, D., Greinacher, A., Kerrigan, S.W. & Watson, S.P. Amplification of bacteria-induced platelet activation is triggered by FcgammaRIIA, integrin alphaIIbbeta3, and platelet factor 4. *Blood* **123**, 3166-74 (2014).
21. Semple, J.W., Italiano, J.E., Jr. & Freedman, J. Platelets and the immune continuum. *Nat Rev Immunol* **11**, 264-74 (2011).
22. Rondina, M.T., Weyrich, A.S. & Zimmerman, G.A. Platelets as cellular effectors of inflammation in vascular diseases. *Circulation research* **112**, 1506-1519 (2013).
23. Thomas, M.R., Wijeyeratne, Y.D., May, J.A., Johnson, A., Heptinstall, S. & Fox, S.C. A platelet P-selectin test predicts adverse cardiovascular events in patients with acute coronary syndromes treated with aspirin and clopidogrel. *Platelets* **25**, 612-618 (2014).
24. McCabe, D.J., Harrison, P., Mackie, I.J., Sidhu, P.S., Purdy, G., Lawrie, A.S., Watt, H., Brown, M.M. & Machin, S.J. Platelet degranulation and monocyte-platelet complex formation are increased in the acute and convalescent phases after ischaemic stroke or transient ischaemic attack. *Br J Haematol* **125**, 777-87 (2004).
25. Cox, D., Kerrigan, S.W. & Watson, S.P. Platelets and the innate immune system: mechanisms of bacterial-induced platelet activation. *Journal of Thrombosis and Haemostasis* **9**, 1097-1107 (2011).
26. Semeraro, N., Ammollo, C.T., Semeraro, F. & Colucci, M. Sepsis, thrombosis and organ dysfunction. *Thromb Res* **129**, 290-5 (2012).
27. Mavrommatis, A.C., Theodoridis, T., Orfanidou, A., Roussos, C., Christopoulou-Kokkinou, V. & Zakynthinos, S. Coagulation system and platelets are fully activated in uncomplicated sepsis. *Crit Care Med* **28**, 451-7 (2000).
28. Rondina, M.T., Carlisle, M., Fraughton, T., Brown, S.M., Miller, R.R., 3rd, Harris, E.S., Weyrich, A.S., Zimmerman, G.A., Supiano, M.A. & Grissom, C.K. Platelet-monocyte aggregate formation and mortality risk in older patients with severe sepsis and septic shock. *J Gerontol A Biol Sci Med Sci* **70**, 225-31 (2015).
29. Russwurm, S., Vickers, J., Meier-Hellmann, A., Spangenberg, P., Bredle, D., Reinhart, K. & Lösche, W. Platelet and leukocyte activation correlate with the severity of septic organ dysfunction. *Shock* **17**, 263-268 (2002).

30. Katz, J.N., Kolappa, K.P. & Becker, R.C. Beyond thrombosis: The versatile platelet in critical illness. *Chest* **139,** 658-668 (2011).
31. Hui, P., Cook, D.J., Lim, W., Fraser, G.A. & Arnold, D.M. The frequency and clinical significance of thrombocytopenia complicating critical illness: A systematic review. *CHEST Journal* **139,** 271-278 (2011).
32. Mandell, L.A., Wunderink, R.G., Anzueto, A., Bartlett, J.G., Campbell, G.D., Dean, N.C., Dowell, S.F., File, T.M., Musher, D.M. & Niederman, M.S. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical infectious diseases* **44,** S27-S72 (2007).
33. Thomas, M.R. & Storey, R.F. Effect of P2Y12 inhibitors on inflammation and immunity. *Thromb Haemost* **114,** 490-7 (2015).
34. Harvey, R.M., Trappetti, C., Mahdi, L.K., Wang, H., McAllister, L.J., Scalvini, A., Paton, A.W. & Paton, J.C. The Variable Region of the Pneumococcal Pathogenicity Island 1 is Responsible for the Unusually High Virulence of a Serotype 1 Isolate. *Infect Immun* (2016).
35. Cremers, A.J., Mobegi, F.M., de Jonge, M.I., van Hijum, S.A., Meis, J.F., Hermans, P.W., Ferwerda, G., Bentley, S.D. & Zomer, A.L. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. *Sci Rep* **5,** 14952 (2015).
36. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30,** 1575-1584 (2002).
37. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559-75 (2007).
38. DeBardeleben, H.K., Lysenko, E.S., Dalia, A.B. & Weiser, J.N. Tolerance of a Phage Element by *Streptococcus pneumoniae* Leads to a Fitness Defect during Colonization. *Journal of Bacteriology* **196,** 2670-2680 (2014).
39. Burghout, P., Bootsma, H.J., Kloosterman, T.G., Bijlsma, J.J.E., de Jongh, C.E., Kuipers, O.P. & Hermans, P.W.M. Search for Genes Essential for Pneumococcal Transformation: the RadA DNA Repair Protein Plays a Role in Genomic Recombination of Donor DNA. *Journal of Bacteriology* **189,** 6540-6550 (2007).
40. Tunjungputri, R.N., Van Der Ven, A.J., Schonsberg, A., Mathan, T.S., Koopmans, P., Roest, M., Fijnheer, R., Groot, P.G. & de Mast, Q. Reduced platelet hyperreactivity and platelet-monocyte aggregation in HIV-infected individuals receiving a raltegravir-based regimen. *Aids* **28,** 2091-6 (2014).

Chapter 8

General discussion

Streptococcus pneumoniae (the pneumococcus) asymptomatically colonizes the nasopharynx of healthy carriers but, can transition into a significant respiratory tract pathogen that is a major cause of global morbidity and mortality. The world health organization estimates that around 1.6 million people die each year worldwide from pneumococcal infections [1]. This burden of invasive pneumococcal disease (IPD) necessitates our constant attention. Imprudent use of antibiotics to manage IPD has led to the development and spread of multidrug resistance, thus demanding a search for novel antimicrobial drugs. Preventive interventions with pneumococcal conjugate vaccines (PCV) have reduced episodes of pneumococcal infections caused by covered vaccine serotypes [2, 3]. However, due to their limited serotype coverage, PCVs also have triggered rapid serotype replacement of vaccine serotypes with non-vaccine serotypes [4, 5]. Recombination events in the capsular loci may give rise to novel 'capsule switched' lineages capable of vaccine-escape [6, 7]. As a result, there is a need for continued surveillance after vaccine introduction to fully understand their impact on pneumococcal populations. Over the past years, our knowledge of high-throughput sequencing technologies and their applications in pneumococcal microbiology has steadily advanced. The aim of this thesis was to further our comprehension of the etiology and epidemiology of *S. pneumoniae* using *in silico* analyses combined with *in vivo* and *in vitro* assays. We performed in-depth studies on genomes of relevant pneumococcal isolates and corresponding clinical metadata to better understand their population genomics, and relate their genetic profiles as well as genotypic diversity with: (i) causation of IPD, (ii) clinical manifestation of pneumococcal disease and (iii) antibiotic resistance. Furthermore, we studied the possibility of finding novel antibiotics that target the pneumococcus as well as *Haemophilus influenzae* and *Moraxella catarrhalis*, which together with the *S. pneumoniae* constitute the major bacterial causes of respiratory tract infections. This chapter summarizes the findings of our studies and discusses their allusions in the light of the current state of literature and future healthcare and research applications.

High throughput discovery of potential drug targets

"A post-antibiotic era means, in effect, an end to modern medicine as we know it." Dr. Margaret Chan [8]

The specter of antibiotic resistance is not a new phenomenon. The founder of antibiotics, Sir Alexander Fleming, had observed that exposing microbes to penicillin concentrations not sufficient to kill them lead to resistance. In his 1945 Nobel Prize winner's lecture [9] he said: "It is not difficult to make microbes resistant to penicillin in the laboratory by exposing them to concentrations not sufficient to kill them, and the same thing has occasionally happened in the body. The time may come when penicillin can be bought by anyone in the shops. Then there is the danger that the ignorant man may easily underdose himself and by exposing his microbes to non-lethal quantities of the drug

make them resistant." True to his word, antimicrobial resistance is inexorably on the rise and it is already threatening most if not all first-line antimicrobials including penicillin [10-13]. The inevitable spread could reverse decades of medical progress and enfeeble healthcare systems globally. Unless action is taken, experts are warning of a "post-antibiotic era [8]" when simple, currently treatable infections, will once again kill unabated. Thus far, antibiotic resistance has been counteracted by administering a combination of antibiotics, and by chemically modifying current antibiotics or isolating derivatives that are less easily degraded or extruded by the bacteria. While antibiotic resistance is common in most pathogenic bacteria, resistant *S. pneumoniae* are particularly rampant that in approximately thirty percent of severe IPD cases in the USA, the bacterium is fully resistant to one or more clinically relevant antibiotics [14]. Additionally, the pneumococcus is a highly recombinogenic and competent bacteria that readily takes up resistance-conferring DNA fragments [15], and can adjust its transformation rates in response to various environmental stimuli, including vaccines and antibiotics [7, 15, 16]. To curb resistance and effectively manage IPD, development of new classes of antimicrobials is necessary. Identification of genes essential for bacterial growth and viability is prerequisite to discovery of molecular targets upon which drugs could subsequently be developed. Next generation sequencing technologies have granted access to higher levels of genome organization making it easier to identify gene function and potentially essential genes.

In **Chapter 3**, we explore a proof-of-concept study for the discovery of essential genes in the pneumococcus which form promising molecular targets upon which antimicrobial drugs could be developed. We employed various *in vitro* and *in silico* analyses to address caveats in traditional drug discovery pipelines that hamper speed and precision. Identification of essential genes is a prerequisite to the discovery of drug target. This process customarily relied on rigorous analysis of fitness defects of single mutants through appropriate *in vitro* or *in vivo* conditions with techniques that were far from high-throughput. Signature Tagged Mutagenesis; STM marked the first step towards high-throughput screening for essential genes [17]. In STM, single mutants in a pool of up to 96 tagged mutants are screened before and after a conditional challenge by detecting mutant-specific DNA tags using southern hybridization. STM, owing to the DNA tag, is highly specific but could only screen a few mutants at a time. The advent of microarray-based methods like Genomic Array Footprinting; GAF and Transposon Site Hybridization; TraSH [18-20] further improved on the throughput but, these methods were prone to low resolution and inaccurate results due to cross-hybridization, impreciseness of locating the insertion site, and inherent drawbacks of the microarray platform such as skewed signals.

With the recently developed technologies like Tn-seq [21], TraDIS [22], INSeq [23], and HITS [24] that rely on high-throughput sequencing of transposon insertion sites, transposon mutant libraries can now be screened at higher resolutions in tracking insertion sites. We exploited an integrative approach that uses genome-wide transposon

mutagenesis, high-throughput transposon sequencing and analysis with the Tn-seq schema [21] implemented in the ESSENTIALS pipeline [25], and comparative genomics; to discover essential genes and prioritize drug targets in certain strains of *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis*: these are the most prevalent bacteria that cause respiratory tract infections; RTI [26]. This approach explores the exhaustive power of transposon insertions sequencing (Tn-seq) complemented with *in silico* tracking to rapidly identify essential genes. Like in previous essentiality screens, it was observed that approximately 20% of all genes in these species are essential to growth and viability. Candidate essential genes were further subjected to a comparative selection criterion to identify a subset of 249 potential target proteins that have no homologues in the human host and human microbiota. This subset constituted of 67 genes that are verified target to 75 FDA-approved antimicrobial and 35 other small molecule inhibitors. The capacity to identify some verity was likely to increase confidence in moving these targets into drug development. In that respect, we experimentally validated the inhibition of bacterial growth using commercially available small-molecule inhibitors of four novel targets using Kirby-Bauer disk diffusion susceptibility test [27]. This approach relied on literature for known inhibitors and assumed that the inhibitor molecule did not target alternative pathways to cause the lethal phenotype. Therefore, it was not unequivocally conclusive on the inhibited function. Nonetheless, the small-scale laboratory assays we performed have provided leads on which analogous active compound could be synthesized and optimized. We further tested for the toxicity of these inhibitors on human cell-lines and computed their effective inhibitory concentrations. This approach addresses issues that hamper accurate and fast discovery of drug targets, and could speed up the process of rational drug design.

High throughput genome sequencing and computational approaches are already used as an addendum to circumvent the tedious laboratory screens conventionally used to screen for essential genes and prioritize drug targets [28-32]. As reviewed in **Chapter 2**, transposon insertion sequencing methods, including Tn-seq, show a remarkable degree of accuracy and sensitivity in predicting gene essentiality. Nonetheless, for these methods to be reliable, they demand saturated mutant libraries and precise deep sequencing making them costly and laborious. This was evident in the number of essential genes we identified in each strain relative to the corresponding transposon mutant libraries. The more saturated libraries of *S. pneumoniae* R6 and TIGR4 strains led to a small but perhaps more accurate set of essential genes; 325 and 414 respectively. These were expected to contain less false positives as compared with *H. influenzae* 86-028NP which had 532 essential genes from a mutant library that was less saturated. Other *ab initio* computational approaches developed to predict essential genes and drug targets are feeble in terms of predicting conditionally essential genes, scaling to different organisms or experimental conditions, and they often require seed biological or biophysicochemical information which is difficult to obtain [33, 34]. Experimental validation of gene function in also expected. This was traditionally achieved by gene-

knockout or gene silencing, and subsequently observing the ensuing phenotype, a process that often takes a lot of time and resources to accomplish. Luckily, there is optimism around the future of computer-guided drug-discovery. Recently developed high-throughput molecular docking (HTMD) approaches have become increasingly important tool for drug discovery [35]. Compared with traditional experimental approaches, HTMD allows for a more direct and rational screening of millions of compounds for leads at low cost and high effectiveness. Because most HTMD methods require structural information of the target proteins and a database of chemicals, we did not perform these assays. We anticipate that future research will take advantage of these high-throughput methods and identify promising clinical candidates that target the essential pathways we have reported.

Overall, advances in high-throughput approaches have permitted a deeper understanding of gene function. This chapter demonstrates that genome sequencing technologies and high-throughput genomic screens could reliably be incorporated into rational drug design to improve on speed, precision, and robustness. Most importantly, we have improved the general understanding of bacterial pathogens of the respiratory tract, and identified novel drug targets and putative inhibitors that form a solid basis for experimental validation. It is our hope that a follow-up studies will lead to the development of novel antimicrobial drugs against major bacterial pathogens of the respiratory tract.

The post-vaccine population genomics of invasive pneumococci

The efficient competence machinery of the pneumococcus enables this microorganism to readily accept exogenous genetic material and evolve rapidly [36, 37]. Strategies to control IPD continue being deployed despite the evident dearth of a detailed understanding of how the pneumococcus evolves in response to these clinical interventions. It is therefore critical to precisely monitor circulating pneumococcal strains and evaluate emerging evolutionary patterns in response to clinical interventions. **Chapter 4** has contributed to our understanding of the post-vaccine population dynamics of invasive pneumococcus in adults through 'herd protection'. This is a form of indirect protection that occurs when a large percentage of a population has been vaccinated. In such a population where the majority is immune to an infection, the chain of infection is disrupted thereby stopping or slowing the spread of disease and providing a measure of protection for the few individuals who are not vaccinated [38]. We elucidated the genetic variations that occurred in sequenced genomes of 350 pneumococcal isolates from adult IPD patients during the years between 2000 and 2011, the period before and after the introduction of 7-valent PCV (PCV7) in the Netherlands: PCV7 was introduced into the Dutch National Immunization program for pediatrics in 2007. The 'core genome' phylogeny; represented by genes present in single copies in all 350 isolates, revealed a tight clustering of clades according to the capsular serotypes, indicating a correlation between genotype and serotype. Like in nasopharyngeal carriage [7], a steady decrease

in adulthood IPD cases caused by vaccine-types was observed after the introduction of PCV7. This reduction in cases of IPD in unimmunized adults following a mass pediatric immunization supports the herd protection (cross-protection) effect of PCV. Also, there was a massive serotype replacement of vaccine-types with non-vaccine serotypes in our disease cohorts. Post-vaccine *S. pneumoniae* infections were dominated by strains expressing non-vaccine capsular types mainly serotypes 1, 3, 7F, and 8, which have surprisingly been reported to be less frequent in asymptomatic carriage [39].

The selective pressure induced by pneumococcal conjugate vaccines has been implicated in causing capsular switches although in rare occasions [7]. Research has, however, reported that the existence of capsular switch singly should not considerably impact the efficacy of PCV on IPD incidence [40]. Nonetheless, recombination of capsular polysaccharide synthesis genes has been recognized as the main driver of structural and antigenic heterogeneity of capsule types. Therefore, the significant PCV-induced recombination may directly lead to vaccine-escape when a vaccine-type takes up and expresses a non-vaccine capsule [6], or trigger the development of chimeric serotypes that may diverge in their interaction with the host immune system [41]. In our unvaccinated adults, capsular switches -typified by the presence of an eccentric serotype within an otherwise monophyletic clade- involved the acquisition of both vaccine and non-vaccine capsular serotypes, and were only found in isolates collected before the introduction of PCV7. This precludes adaptation to vaccine-pressure as the driver of these capsule switches.

The introduction of PCV7 also caused a sharp drop in the diversity the core and accessory genome shortly after 2007. However, subsequent years saw a considerable re-expansion of the genomes and a return towards equilibrium which was mainly contributed by genes extant pre-vaccination. This tendency of the strains to recover the genes after perturbation by vaccines perhaps points to the importance of these genes to the pneumococci for thriving during carriage and invasive disease. Alternatively, strains carrying these genes may have previously existed, albeit unsuccessful in outcompeting the vaccine-types in the absence of vaccine pressure. This dynamic post-vaccine reshuffling of the pneumococcal genome could actually enlarge the repertoire of virulence factors subsequently altering strains invasiveness potential. For example, Cremers *et al* [42] observed in their IPD cohort that patients infected by pneumococcal strains expressing zinc metalloproteinase C (ZmpC) in their genome often presented more severe clinical symptoms and more frequently required intensive care as compared to patients infected with stains lacking *zmpC*. These findings demonstrate that presence of certain pneumococcal genetic factors may therefore be useful in monitoring the clinical manifestation and predict predicting the course of IPD.

The rollout of massive immunization programs with PCV around the world has remarkably reduced episodes of IPD caused by vaccine types both in vaccinated

pediatrics and in unvaccinated adults through herd protection. However, the full effects of mass vaccination with PCVs are not well understood. Studies have reported a risk of the emergence of replacement invasive vaccine-escape strains. Continuous surveillance is therefore necessary to better understand the effects of mass vaccination on circulating pneumococcal strains. This study revealed the herd immunity effect and the perturbations of pediatric immunization with PCV7 on the population epidemiology of invasive pneumococci in adults. Furthermore, the chapter presents whole-genome sequencing as a powerful tool for molecular epidemiological surveillance of pneumococcal strains at a genomic level resolution. These advances will significantly influence the modelling of future interventions.

Correlating pneumococcal genetic profiles with phenotypes of antibiotic resistance and clinical manifestation of IPD

Because the capsule is known to be vital for virulence and is the target for the PCV-induced serotype-specific protection against IPD, there has been little focus on other genotypic factors in development of vaccines. Until now, the pneumococcal invasiveness has been entirely based on the capsular serotypes. However, it is highly unlikely that capsular serotype solitarily outlines pneumococcal pathogenicity [43]. If this was the case, genetically divergent strains expressing identical capsular serotype should have a similar potential for invasiveness. Other genetic factors, but the capsule, have been shown to determine variations in pneumococcal pathogenicity [43-46]. Therefore, sequence types will perhaps be insufficient to explain the invasiveness of individual strains especially within a highly recombinogenic species such as the pneumococcus. Whole-genome sequencing holds the key to understanding the interplay between ecology and genetic adaptation, and identifying other genes outside the capsular loci, or genetic variations that confer invasiveness [46, 47].

Chapters 5 and 6 explore the use of whole-genome sequencing and genome-wide association studies (GWAS) to give insight into how pneumococcal genotypic variants give rise to antimicrobial resistance and clinical eventuality of invasive pneumococcal disease respectively. Research has strived in identifying variations, especially single nucleotide polymorphisms (SNP) and presence or absence of certain genes, in genomes of bacterial pathogens - including the pneumococcus - that confer resistance to various antibiotics [48-50]. In a world in which antibiotics are widely used, we need to know much more in general about development of inexorable antibiotic resistance and the emergence of multi-drug resistant invasive strains of *S. pneumoniae*, to help improve diagnosis and choice of antibiotic(s) [51]. A recently published study explores the use of de Bruijn graphs on genome sequence data to identify species and detect antibiotic resistance profiles in clinical samples of *Staphylococcus aureus* and *Mycobacterium tuberculosis in silico* [48]. The model showed over 99% sensitivity and specificity across 12 antibiotics for *S. aureus,* and a slightly lower specificity (98.5%) at 82.6% sensitivity for *M.*

tuberculosis: these values are comparable to gold-standard phenotypic methods, demonstrating the feasibility of using genome sequencing data with *in silico* approaches in rapid diagnostics.

In **Chapter 5**, we investigated the distance to antimicrobial resistance using 1682 literature and in-house *S. pneumoniae* genomes. First we identified resistance-conferring SNPs using GWAS. The clonal nature of pneumococcal lineages and their high rates of recombination demand a correction for population stratification to break the strong linkage disequilibrium which often inflates the association test *p*-values obscuring true causal variants among 'hitch-hikers' in association studies. Population stratification can be performed using genomic control [52]; whereby all *p*-values are normalized with a single inflation factor λ or through inference of ancestry by identifying genetic subpopulations within the overall population and testing for associations conditional on these subpopulations. Population substructures are heuristically inferred using various approached, for example Bayesian Analysis of Population Structure; BAPS [53], principal component analysis; PCA in EIGENSTRAT [54] and multi-dimensional scaling in PLINK [55]. Starting with a concatenated super-alignment of all variant regions on the core genome of 1682 pneumococcal isolates, we determined subpopulations as sequence clusters; SCs using BAPS. Also, we generated a phylogeny on the alignment and inferred SCs using a clustering method proposed by Prosperi *et al* [56] which partitions phylogenies. We tested for associations between whole-genome SNPs and resistance to antibiotics (penicillin, trimethoprim, co-trimoxazole, erythromycin, ofloxacin, ciprofloxacin, and tobramycin) conditional on these SCs. Resistance-conferring mutations (SNPs) were localized mainly in genes previously associated with resistance against the classes of antibiotics we tested. They include penicillin binding proteins (*pbp1A, pbp2A*, and *pbpX*); pneumococcal surface protein (*pspA, pspC*); multi-drug resistance efflux pump (*pmrA*); dihydrofolate reductase (*dyr*); and dihydropteroate synthase (*folP/sulA*). Other mutations however appeared in novel putative resistance genes like transcription factor (*rarA*) and heme exporter protein A (*ccmA*) for fluoroquinolone resistance, and the cell division proteins (*gpsB*, *ftsL*, *ftsX*) for penicillin resistance.

A plasmid-borne efflux pump *oqxAB* in *Escherichia coli* confers resistance to quinolones [57]. More recently studies have shown that mutations in an *oqxAB* homolog in *Klebsiella pneumoniae* (*oqxR*) led to overexpression of *rarA,* which encodes the *oqxAB* transcriptional regulator, eliciting cross-resistance to quinolones [58-61]. RarA may therefore be implicated in up-regulation or down-regulation of the expression of the *oqxAB* drug efflux pumps. More studies will be required to characterize the exact mechanisms, whether direct or indirect, through which the mutations seen on *rarA* influence quinolone resistance in the pneumococcus. While dissemination of conjugative plasmid-derived antibiotic resistance genes is common in bacteria [62-64], to our knowledge, mutational-driven resistance by plasmid-borne genes has not been reported yet. *CcmA* on the other hand is implicated in biofilm formation in *Streptococcus gordonii*.

Loss of expression could have a variety of effects on metabolic activity, and on protein and peptide secretion and trafficking [65]. Moreover, *ccmA* mutants show a phenotype of aberrant cell wall morphology [66], indicating that this gene may be involved in cell-wall biogenesis. Biofilm formation is a simple strategy of microbial survival and persistence. The biofilm formation process itself is complex, involving a variety of genes, and influenced by several environmental factors [67, 68]. Bacteria in biofilm communities exhibit an inherent tolerance to antibiotics and host immunity [69]. Biofilm formation in pneumococci has been shown to enhance strains fitness and virulence [67, 68]. Of particular interest is the ability of antibiotic-susceptible strains to form thicker biofilms compared to resistant strains [70], perhaps as a strategy to resist antibiotics. The functions played by *ccmA* in promoting bacterial sessility may be suggestive of its importance in antibiotic resistance. Donati et al [71] predicted a fast-growing pneumococcal pan-genome characterized by a large gene repertoire, and postulated that the microorganism remains open to novel evolutionary possibilities by rapidly acquiring and integrating novel genes.

In recent years, creating antimicrobial compounds designed to inhibit the bacterial divisome has been recognized as a promising strategy for antibiotic attack [72, 73]. Interfering with proteins involved in bacterial cell division blocks the pathogen's proliferation leading death. *GpsB* and *ftsL* play a role in cell division and are essential for complete cell wall formation. *FtsL* forms a complex with other essential genes that co-localize at the division site to regulate the cell-division [74]; deletion of one or more components destabilizes this complex hampering the cell division apparatus [74-76]. On the other hand, mutants of *GpsB* show phenotypes of cell deformation similar to those observed in methicillin-mediated inhibition of the penicillin binding protein (*Pbp2x*) [77]. Whether these cell division proteins potentially interact with *pbp* genes, either directly or indirectly through regulation or participation in cell wall formation remains unclear. It will be interesting and important for future studies to experimentally elucidate and characterize the mechanisms of how alterations in transcription regulation, biofilm formation, and cell division factors influence resistance to quinolone and beta-lactams respectively. All in all, by aggregating the odds ratios (ORs) effect of candidate SNPs for each genome, we were able to partition the resistance phenotypes and identify resistant strains in each population.

Whole-genome sequencing technologies have great potential, in the future, in simplifying the diagnostic laboratory workflow thus becoming a part of routine clinical practice. Currently, the phenotypic tests used to determine antibiograms take days to complete. There is a shift towards rapid sequencing-based diagnostics, but the link between genotype and phenotype has not been perfected yet. Studies like this one should results into a model that allows the prediction of antibiograms from genome sequencing data. It is also noteworthy that the NCBI has started a database; National Database of Antibiotic Resistant Organisms (NDARO), which collates and makes publicly available the resistance

phenotypes and corresponding genome sequences of infectious organisms. Such initiatives will provide a unique platform for rapid assessment of the presence of resistance genes using assembled genomes. At present however, sequencing-based diagnostics require optimizations to eliminate methodological and technical errors from sample collection, sequencing and analysis because these generally lead to false positive results. Nonetheless, these technologies have aided in understanding the evolutionary variations and observing their dynamics in evolving pneumococcal lineages. This study shows that whole-genome sequencing could be employed to develop novel easy-to-use clinical diagnostic assays that may provide consistent results in rapid detection of drug-resistant pneumococcal strains. Moreover, the data generated here is invaluable for evaluating pneumococcal selection to antibiotic resistance and evolution of strains in response to antimicrobial drugs. This information could inform future drug development and enhance surveillance of circulating pneumococcal strains to be vigilant on the emergence of novel resistance mechanisms.

Chapter 6 improves our understanding on how pneumococcal genotypic variations affect the severity of IPD. The utilization of whole-genome sequencing and analysis technologies to predict and characterize pathogen virulence has been emphasized in recent times [78, 79]. To discover genetic variants that associate with clinical manifestation of IPD, we analyzed the accessory genome (the flexible gene pool) of 350 invasive pneumococci isolated from patients admitted with pneumococcal disease in two hospitals in Nijmegen, the Netherlands. Generally, IPD patients exhibit a plethora of diverse clinical outcomes, begging the question whether certain pneumococcal genotypic traits contribute to these outcomes. If indeed there are essential pneumococcal traits linked to disease manifestation, they have important implications for prevention of IPD. In this study, we identified genes significantly associated with pneumonia, meningitis, and 30-day mortality. They included genes previously associated with these clinical phenotypes and novel genes that required experimental validation. Extensive vaccination with current vaccines which target a limited number of circulating pneumococcal strains has led to significant serotype replacement that rescinds the benefits of pneumococcal vaccination. The possibility still exists of identifying proteins conferring invasiveness or contributing to clinical eventuality of disease, and using them as candidate targets for future vaccine. In this study, most of the disease-associated factors we identified changed in percentage prevalence following the introduction of PCV7. They include CDP-Glycerol-1-Phosphate Biosynthetic Protein (*gct*) and phospholipase A2 (*slaA*) that have been implicated in meningitis and *pblB* which is implicated in promoting mortality. It is still unclear whether these perturbations affect virulence factors and consequently the eventuality of pneumococcal disease. However, research has shown that some genotypic factors associated with the replacing non-vaccine strains might play a role in exacerbating the clinical manifestation of pneumococcal disease [42] and promoting antibiotic resistance [80]. Unfortunately, we lacked a second cohort with relevant metadata to validate our findings and increase confidence in verity. It is our hope that as the costs for

genome-sequencing decreases, more data will be made available for similar studies. Overall, our results suggest that modeling of future pneumococcal vaccines requires diversification from exclusively targeting the capsular serotypes and capitalize on tailoring for other disease inflating genetic factors in circulating pneumococcal strains. By increasing the range of vaccine targets, the selective pressure on the pneumococcal populations in general could be reduced, subsequently limiting the undesirable serotype replacement and capsular switches thus enhancing the efficacy of future vaccines. The findings in this chapter also recap the biological significance of genetic factors that lie beneath the capsule since the replacement strains may exhibit different virulence potential. The study also reinforces that perhaps more studies need to be performed in this realm to monitor the emerging genotypic and phenotypic patterns following clinical interventions.

Chapter 7 further explores the clinical role of a phage-encoded gene *pblB* (OG_ 17) in the mortality of patients with IPD. The majority of clinically relevant pneumococcal isolates have been shown to carry bacteriophages [81]. It is also widely accepted that bacteria could acquire virulence properties from these mobile genetic elements [82]; however, the role of bacteriophages in pathogenesis of IPD remains largely unknown. A GWAS analysis of 350 pneumococcal genomes revealed that *pblB* was positively associated (*p*-value = 0.00034 Bonferroni corrected for multiple testing) with 'all-cause' mortality in IPD patients within the first 30 days of hospitalization (30-day mortality). PblB, phage-encoded surface protein, was shown to partly mediate *in vitro* platelets binding by *Streptococcus mitis* [83, 84]. More recent studies on pneumophages have shown the importance of *pblB* in promoting persistence and fitness both in the nasopharynx and the lungs, and in facilitating pneumococcal adhesion to platelets and epithelial cells [85, 86]. The adhesion is thought to promote pneumococcal clonal success and pathogenicity. For instance, an *in vivo* mouse model showed that *pblB* deletion resulted in reduced adherence, biofilm formation, reduced initial infection within the lungs, and a reduction in the number of circulating platelets [87].

We observed an inclination towards later fatality among IPD cases involving pneumococci carrying the *pblB* gene; *pblB*+ (mean 6.6 vs 3.0 days, *p*-value 0.09). Even though expression of prophage genes was known to be induced by antibiotics, it was unknown whether pneumophages could specifically be induced by fluoroquinolones. In our study sub-lethal doses of ciprofloxacin and levofloxacin were optimal for the induction of *pblB*, whereas beta-lactam antibiotic penicillin G did not induce the expression. However, there were no differences in 30-day mortality between patients who were treated with fluoroquinolones and those who were not. Perhaps a larger cohort, preferably from a different country where fluoroquinolones are used to manage IPD, would be necessary to verify these findings. Platelet activation occurs in patients with pneumonia or sepsis and has previously been shown to enhance development of hyperinflammation, microthrombosis, myocardial infarction and multiple organ failure [88, 89]. Increased

platelet activation has also been shown to be a predictor of mortality in older septic patients [90]. In our study after the induction of *pblB* with levofloxacin, the pneumococci triggered higher platelet surface expression of P-selectin and platelet monocyte complex formation. This study highlights the clinical utility of bacterial GWAS, followed by functional characterization in identifying bacterial factors involved in disease severity.

Concluding remarks

The next-generation genome sequencing (NGS) and analysis technologies such as metagenomics-based screening are very likely to become an alternative to the conventional laboratory methods used for bacteriology. In fact, the technology is already being used to complement traditional experimental approaches for molecular diagnostics, rational vaccine and antibiotic design, molecular epidemiology and mapping of disease outbreaks [91-93]. In addition, NGS is likely to transform our insight on mechanisms of infection, sanctioning rational preventive measures [94] thus reducing the risk of unwarranted interventions [95]. NGS and metagenomics-based screens have also allowed for a means to circumvent culturing steps, and determine the presence of particular 'risky' functionalities, including virulence factors and antibiotic resistance markers, collectively transforming the way we do microbial research. Presently however, the different NGS techniques have different attributes impeding the development of a universal methodology for data management and analysis. Software developers are also facing challenges to cope with the escalating amount of genomes and genome-sequencing data, and ensure harmonized integration with relevant metadata as well as prudent stewardship and sharing. Additionally NGS is not cost-effective for routine and offers only predictions: phenotypic validations might still be necessary. Nonetheless, efforts are currently being made to create a universal database, dubbed *global pneumococcal sequencing project*; GPS that collates pneumococcal isolates, their genome sequences, and the associated clinical metadata [96]. In addition, the National Institutes of Health has financed, to a tune of $10 million, a multidisciplinary team of researchers to study the role of the immune system in the emergence of antibiotic-resistant bacteria. The team will apply cutting-edge high-throughput genomics applications and methods to elucidate the interplay between bacterial genetic adaptation, host immunity and clinical intervention. Already, genome sequencing and bioinformatics have proven to be pivotal in these ventures and they have provided critical insights into the biology of the pneumococcus. When completed, these studies will provide platforms for observing the genomic diversity and genetic adaptations of *S. pneumoniae* strains in response to current clinical interventions and the host immune response. The aim is to guide modelling of better next-generation interventions that avoid escape and resistance. Most of the data analyzed in this thesis will be integrated into the global database adding new data-points to the repository.

Largely, the findings presented in this thesis have broadened our understanding in the utilization of whole-genome sequencing and bioinformatics in rational designing of antimicrobial drugs, assessment of the drug resistance profiles in strains and designing of novel diagnostic assays; monitoring the dynamics of population genomics and adaptation to clinical intervention; characterizing pneumococcal virulence repertoires; and surveying the acquisition and/or dissemination of virulence factors to provide a future trajectory of the invasiveness potential of circulating strains. Our methods and findings also dovetail neatly with current technological and literature advances, and they open up new avenues for future pneumococcal research. As such, this thesis puts to the forefront the important practical applications of bioinformatics and NGS in augmenting and furthering pneumococcal bacteriology.

References

1. World Health Organization. Pneumococcal vaccines. *Weekly epidemiological record* **78,** 110-119 (2003).
2. Bogaert, D., Hermans, P.W.M., Adrian, P.V., Rümke, H.C. & de Groot, R. Pneumococcal vaccines: an update on current strategies. *Vaccine* **22,** 2209-2220 (2004).
3. van Deursen, A.M., van Mens, S.P., Sanders, E.A., Vlaminckx, B.J., de Melker, H.E., Schouls, L.M., de Greeff, S.C. & van der Ende, A. Invasive pneumococcal disease and 7-valent pneumococcal conjugate vaccine, the Netherlands. *Emerg Infect Dis* **18,** 1729-37 (2012).
4. Hanage, W.P., Finkelstein, J.A., Huang, S.S., Pelton, S.I., Stevenson, A.E., Kleinman, K., Hinrichsen, V.L. & Fraser, C. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics* **2,** 80-84 (2010).
5. Cremers, A.J.H., Mobegi, F.M., de Jonge, M.I., van Hijum, S.A.F.T., Meis, J.F., Hermans, P.W.M., Ferwerda, G., Bentley, S.D. & Zomer, A.L. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. *Scientific Reports* **5,** 14952 (2015).
6. Brueggemann, A.B., Pai, R., Crook, D.W. & Beall, B. Vaccine Escape Recombinants Emerge after Pneumococcal Vaccination in the United States. *PLoS Pathog* **3,** e168 (2007).
7. Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P. & Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45,** 656-63 (2013).
8. Dr Margaret Chan. Antimicrobial resistance in the European Union and the world: *The EU's contributions to the solutions of the global antimicrobial resistance problem*. in *Keynote address at the Conference on Combating Antimicrobial Resistance: time for action (14 March 2012)* (Copenhagen, Denmark, 2012).
9. Fleming, A., Chain, E.B. & Florey, H. Penicillin: Nobel Lecture, December 11, 1945. Vol. 2016 (The Nobel Foundation, Stockholm, Sweden, 1945).
10. Alvares, J.R., Mantese, O.C., Paula, A.d., Wolkers, P.C.B., Almeida, V.V.P., Almeida, S.C.G., Guerra, M.L.L.S. & Brandileone, M.C.d.C. Prevalence of pneumococcal serotypes and resistance to antimicrobial agents in patients with meningitis: ten-year analysis. *Brazilian Journal of Infectious Diseases* **15,** 22-27 (2011).
11. Davidson, R., Cavalcanti, R., Brunton, J.L., Bast, D.J., de Azavedo, J.C.S., Kibsey, P., Fleming, C. & Low, D.E. Resistance to Levofloxacin and Failure of Treatment of Pneumococcal Pneumonia. *New England Journal of Medicine* **346,** 747-750 (2002).
12. Descheemaeker, P., Chapelle, S., Lammens, C., Hauchecorne, M., Wijdooghe, M., Vandamme, P., Ieven, M. & Goossens, H. Macrolide resistance and erythromycin resistance determinants among Belgian *Streptococcus pyogenes* and *Streptococcus pneumoniae* isolates. *Journal of Antimicrobial Chemotherapy* **45,** 167-173 (2000).
13. Feikin, D.R., Schuchat, A., Kolczak, M., Barrett, N.L., Harrison, L.H., Lefkowitz, L., McGeer, A., Farley, M.M., Vugia, D.J., Lexau, C., Stefonek, K.R., Patterson, J.E. & Jorgensen, J.H. Mortality from invasive pneumococcal pneumonia in the era of antibiotic resistance, 1995-1997. *Am J Public Health* **90,** 223-9 (2000).
14. U.S. Centers for Disease Control and Prevention. Antibiotic Resistance Threats in the United States, 2013. (2014).

15. Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P. & Bentley, S.D. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46,** 305-9 (2014).
16. Evans, B.A. & Rozen, D.E. Significant variation in transformation frequency in *Streptococcus pneumoniae*. *The ISME journal* **7,** 791-799 (2013).
17. Hensel, M., Shea, J., Gleeson, C., Jones, M., Dalton, E. & Holden, D. Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269,** 400-403 (1995).
18. Sassetti, C.M., Boyd, D.H. & Rubin, E.J. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci U S A* **98,** 12712-7 (2001).
19. Bijlsma, J.J., Burghout, P., Kloosterman, T.G., Bootsma, H.J., de Jong, A., Hermans, P.W. & Kuipers, O.P. Development of genomic array footprinting for identification of conditionally essential genes in *Streptococcus pneumoniae*. *Appl Environ Microbiol* **73,** 1514-24 (2007).
20. Burghout, P., Bootsma, H.J., Kloosterman, T.G., Bijlsma, J.J., de Jongh, C.E., Kuipers, O.P. & Hermans, P.W. Search for genes essential for pneumococcal transformation: the RADA DNA repair protein plays a role in genomic recombination of donor DNA. *J Bacteriol* **189,** 6540-50 (2007).
21. van Opijnen, T., Bodi, K.L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6,** 767-72 (2009).
22. Langridge, G.C., Phan, M.D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., Wain, J., Parkhill, J. & Turner, A.K. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* **19,** 2308-16 (2009).
23. Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R. & Gordon, J.I. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6,** 279-89 (2009).
24. Gawronski, J.D., Wong, S.M., Giannoukos, G., Ward, D.V. & Akerley, B.J. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A* **106,** 16422-7 (2009).
25. Zomer, A., Burghout, P., Bootsma, H.J., Hermans, P.W. & van Hijum, S.A. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One* **7,** e43012 (2012).
26. Hoban, D.J., Doern, G.V., Fluit, A.C., Roussel-Delvallez, M. & Jones, R.N. Worldwide prevalence of antimicrobial resistance in *Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis* in the SENTRY Antimicrobial Surveillance Program, 1997-1999. *Clin Infect Dis* **32 Suppl 2,** S81-93 (2001).
27. Bauer, A.W., Perry, D.M. & Kirby, W.M. Single-disk antibiotic-sensitivity testing of staphylococci: An analysis of technique and results. *A.M.A. Archives of Internal Medicine* **104,** 208-216 (1959).
28. Paul, M.L.S., Kaur, A., Geete, A. & Sobhia, M.E. Essential gene identification and drug target prioritization in Leishmania species. *Molecular BioSystems* **10,** 1184-1195 (2014).
29. Muzzi, A., Masignani, V. & Rappuoli, R. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discovery Today* **12,** 429-439 (2007).

30. Moule, M.G., Hemsley, C.M., Seet, Q., Guerra-Assunção, J.A., Lim, J., Sarkar-Tyson, M., Clark, T.G., Tan, P.B.O., Titball, R.W., Cuccui, J. & Wren, B.W. Genome-Wide Saturation Mutagenesis of Burkholderia pseudomallei K96243 Predicts Essential Genes and Novel Targets for Antimicrobial Development. *mBio* **5**(2014).
31. Kitano, H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov* **5**, 202-210 (2007).
32. Duffield, M., Cooper, I., McAlister, E., Bayliss, M., Ford, D. & Oyston, P. Predicting conserved essential genes in bacteria: *in silico* identification of putative drug targets. *Mol Biosyst* **6**, 2482-9 (2010).
33. Wang, Y.-Y., Nacher, J.C. & Zhao, X.-M. Predicting drug targets based on protein domains. *Molecular BioSystems* **8**, 1528-1534 (2012).
34. Singh, N.K., Selvam, S.M. & Chakravarthy, P. T-iDT: Tool for Identification of Drug Target in Bacteria and Validation by Mycobacterium Tuberculosis. *In Silico Biology* **6**, 485-493 (2006).
35. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A powerful approach for structure-based drug discovery. *Current computer-aided drug design* **7**, 146-157 (2011).
36. Croucher, N.J., Coupland, P.G., Stevenson, A.E., Callendrello, A., Bentley, S.D. & Hanage, W.P. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* **5**, 5471 (2014).
37. Hanage, W.P., Fraser, C. & Spratt, B.G. The impact of homologous recombination on the generation of diversity in bacteria. *Journal of Theoretical Biology* **239**, 210-219 (2006).
38. Fine, P., Eames, K. & Heymann, D.L. "Herd Immunity": A Rough Guide. *Clinical Infectious Diseases* **52**, 911-916 (2011).
39. Brueggemann, A.B., Griffiths, D.T., Meats, E., Peto, T., Crook, D.W. & Spratt, B.G. Clonal Relationships between Invasive and Carriage *Streptococcus pneumoniae* and Serotype- and Clone-Specific Differences in Invasive Disease Potential. *Journal of Infectious Diseases* **187**, 1424-1432 (2003).
40. Temime, L., Boelle, P.-Y., Opatowski, L. & Guillemot, D. Impact of Capsular Switch on Invasive Pneumococcal Disease Incidence in a Vaccinated Population. *PLoS ONE* **3**, e3244 (2008).
41. Oliver, M.B., van der Linden, M.P.G., Küntzel, S.A., Saad, J.S. & Nahm, M.H. Discovery of *Streptococcus pneumoniae* serotype 6 variants with glycosyltransferases synthesizing two differing repeating units. *Journal of Biological Chemistry* (2013).
42. Cremers, A.J., Kokmeijer, I., Groh, L., de Jonge, M.I. & Ferwerda, G. The role of ZmpC in the clinical manifestation of invasive pneumococcal disease. *Int J Med Microbiol* **304**, 984-9 (2014).
43. Klugman, K.P., Bentley, S.D. & McGee, L. Determinants of Invasiveness Beneath the Capsule of the Pneumococcus. *Journal of Infectious Diseases* **209**, 321-322 (2014).
44. Browall, S., Norman, M., Tångrot, J., Galanis, I., Sjöström, K., Dagerhamn, J., Hellberg, C., Pathak, A., Spadafina, T., Sandgren, A., Bättig, P., Franzén, O., Andersson, B., Örtqvist, Å., Normark, S. & Henriques-Normark, B. Intraclonal Variations Among *Streptococcus pneumoniae* Isolates Influence the Likelihood of Invasive Disease in Children. *Journal of Infectious Diseases* **209**, 377-388 (2014).
45. Sá-Leão, R., Pinto, F., Aguiar, S., Nunes, S., Carriço, J.A., Frazão, N., Gonçalves-Sousa, N., Melo-Cristino, J., de Lencastre, H. & Ramirez, M. Analysis of Invasiveness of Pneumococcal Serotypes and Clones Circulating in Portugal before Widespread Use of

Conjugate Vaccines Reveals Heterogeneous Behavior of Clones Expressing the Same Serotype. *Journal of Clinical Microbiology* **49,** 1369-1375 (2011).
46. Croucher, N.J., Harris, S.R., Fraser, C., Quail, M.A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J.H., Ko, K.S., Pichon, B., Baker, S., Parry, C.M., Lambertsen, L.M., Shahinas, D., Pillai, D.R., Mitchell, T.J., Dougan, G., Tomasz, A., Klugman, K.P., Parkhill, J., Hanage, W.P. & Bentley, S.D. Rapid pneumococcal evolution in response to clinical interventions. *Science (New York, N.Y.)* **331,** 430-434 (2011).
47. Dagerhamn, J., Blomberg, C., Browall, S., Sjöström, K., Morfeldt, E. & Henriques-Normark, B. Determination of Accessory Gene Patterns Predicts the Same Relatedness among Strains of *Streptococcus pneumoniae* as Sequencing of Housekeeping Genes Does and Represents a Novel Approach in Molecular Epidemiology. *Journal of Clinical Microbiology* **46,** 863-868 (2008).
48. Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., de Cesare, M., Piazza, P., Votintseva, A.A., Golubchik, T., Wilson, D.J., Wyllie, D.H., Diel, R., Niemann, S., Feuerriegel, S., Kohl, T.A., Ismail, N., Omar, S.V., Smith, E.G., Buck, D., McVean, G., Walker, A.S., Peto, T.E., Crook, D.W. & Iqbal, Z. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and Mycobacterium tuberculosis. *Nat Commun* **6,** 10063 (2015).
49. Buwembo, W., Aery, S., Rwenyonyi, C.M., Swedberg, G. & Kironde, F. Point Mutations in the folP Gene Partly Explain Sulfonamide Resistance of *Streptococcus mutans*. *Int J Microbiol* **2013,** 367021 (2013).
50. Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E., Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., Turner, P., Bentley, S.D. & Parkhill, J. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet* **10,** e1004547 (2014).
51. Punina, N.V., Makridakis, N.M., Remnev, M.A. & Topunov, A.F. Whole-genome sequencing targets drug-resistant bacterial infections. *Human Genomics* **9,** 1-20 (2015).
52. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997-1004 (1999).
53. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30,** 1224-8 (2013).
54. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38,** 904-909 (2006).
55. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559-75 (2007).
56. Prosperi, M.C.F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M.C., Gismondo, M.R., Sighinolfi, L., Gray, R.R., Salemi, M., Zazzi, M. & De Luca, A. A novel methodology for large-scale phylogeny partition. *Nat Commun* **2,** 321 (2011).
57. Hansen, L.H., Johannesen, E., Burmølle, M., Sørensen, A.H. & Sørensen, S.J. Plasmid-Encoded Multidrug Efflux Pump Conferring Resistance to Olaquindox in Escherichia coli. *Antimicrobial Agents and Chemotherapy* **48,** 3332-3337 (2004).

58. Veleba, M., Higgins, P.G., Gonzalez, G., Seifert, H. & Schneiders, T. Characterization of RarA, a Novel AraC Family Multidrug Resistance Regulator in Klebsiella pneumoniae. *Antimicrobial Agents and Chemotherapy* **56**, 4450-4458 (2012).
59. Bialek-Davenet, S., Lavigne, J.-P., Guyot, K., Mayer, N., Tournebize, R., Brisse, S., Leflon-Guibout, V. & Nicolas-Chanoine, M.-H. Differential contribution of AcrAB and OqxAB efflux pumps to multidrug resistance and virulence in Klebsiella pneumoniae. *Journal of Antimicrobial Chemotherapy* **70**, 81-88 (2015).
60. Rodríguez-Martínez, J.M., Díaz de Alba, P., Briales, A., Machuca, J., Lossa, M., Fernández-Cuenca, F., Rodríguez Baño, J., Martínez-Martínez, L. & Pascual, Á. Contribution of OqxAB efflux pumps to quinolone resistance in extended-spectrum-β-lactamase-producing Klebsiella pneumoniae. *Journal of Antimicrobial Chemotherapy* **68**, 68-73 (2013).
61. Wong, M.H.Y., Chan, E.W.C. & Chen, S. Evolution and Dissemination of OqxAB-Like Efflux Pumps, an Emerging Quinolone Resistance Determinant among Members of Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy* **59**, 3290-3297 (2015).
62. Zhao, J., Chen, Z., Chen, S., Deng, Y., Liu, Y., Tian, W., Huang, X., Wu, C., Sun, Y., Sun, Y., Zeng, Z. & Liu, J.-H. Prevalence and Dissemination of oqxAB in Escherichia coli Isolates from Animals, Farmworkers, and the Environment. *Antimicrobial Agents and Chemotherapy* **54**, 4219-4224 (2010).
63. Wong, M.H.Y. & Chen, S. First Detection of oqxAB in Salmonella spp. Isolated from Food. *Antimicrobial Agents and Chemotherapy* **57**, 658-660 (2013).
64. Kim, H.B., Wang, M., Park, C.H., Kim, E.-C., Jacoby, G.A. & Hooper, D.C. oqxAB Encoding a Multidrug Efflux Pump in Human Clinical Isolates of Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy* **53**, 3582-3584 (2009).
65. Kuboniwa, M., Tribble, G.D., James, C.E., Kilic, A.O., Tao, L., Herzberg, M.C., Shizukuishi, S. & Lamont, R.J. *Streptococcus gordonii* utilizes several distinct gene functions to recruit Porphyromonas gingivalis into a mixed community. *Molecular Microbiology* **60**, 121-139 (2006).
66. Hay, N.A., Tipper, D.J., Gygi, D. & Hughes, C. A Novel Membrane Protein Influencing Cell Shape and Multicellular Swarming of Proteus mirabilis. *Journal of Bacteriology* **181**, 2008-2016 (1999).
67. Moscoso, M., García, E. & López, R. Biofilm Formation by *Streptococcus pneumoniae*: Role of Choline, Extracellular DNA, and Capsular Polysaccharide in Microbial Accretion. *Journal of Bacteriology* **188**, 7785-7795 (2006).
68. Domenech, M., García, E. & Moscoso, M. Biofilm formation in *Streptococcus pneumoniae*. *Microbial biotechnology* **5**, 455-465 (2012).
69. Parsek, M.R. & Singh, P.K. Bacterial Biofilms: An Emerging Link to Disease Pathogenesis. *Annual Review of Microbiology* **57**, 677-701 (2003).
70. Camilli, R., Pantosti, A. & Baldassarri, L. Contribution of serotype and genetic background to biofilm formation by *Streptococcus pneumoniae*. *European Journal of Clinical Microbiology & Infectious Diseases* **30**, 97-102 (2010).
71. Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., Oggioni, M., Dunning Hotopp, J.C., Hu, F.Z., Riley, D.R., Covacci, A., Mitchell, T.J., Bentley, S.D., Kilian, M., Ehrlich, G.D., Rappuoli, R., Moxon, E.R. & Masignani, V. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biology* **11**, R107-R107 (2010).
72. Sass, P. & Brötz-Oesterhelt, H. Bacterial cell division as a target for new antibiotics. *Current Opinion in Microbiology* **16**, 522-530 (2013).

73. Lock, R.L. & Harry, E.J. Cell-division inhibitors: new insights for future antibiotics. *Nat Rev Drug Discov* **7**, 324-338 (2008).
74. Noirclerc-Savoye, M., Le Gouëllec, A., Morlot, C., Dideberg, O., Vernet, T. & Zapun, A. *In vitro* reconstitution of a trimeric complex of DivIB, DivIC and FtsL, and their transient co-localization at the division site in *Streptococcus pneumoniae*. *Molecular Microbiology* **55**, 413-424 (2005).
75. Daniel, R.A., Harry, E.J., Katis, V.L., Wake, R.G. & Errington, J. Characterization of the essential cell division gene ftsL (yllD) of Bacillus subtilis and its role in the assembly of the division apparatus. *Molecular Microbiology* **29**, 593-604 (1998).
76. Weiss, D.S., Chen, J.C., Ghigo, J.-M., Boyd, D. & Beckwith, J. Localization of FtsI (PBP3) to the Septal Ring Requires Its Membrane Anchor, the Z Ring, FtsA, FtsQ, and FtsL. *Journal of Bacteriology* **181**, 508-520 (1999).
77. Land, A.D., Tsui, H.-C.T., Kocaoglu, O., Vella, S.A., Shaw, S.L., Keen, S.K., Sham, L.-T., Carlson, E.E. & Winkler, M.E. Requirement of Essential Pbp2x and GpsB for Septal Ring Closure in *Streptococcus pneumoniae* D39. *Molecular microbiology* **90**, 939-955 (2013).
78. Grad, Y.H. & Lipsitch, M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol* **15**, 538 (2014).
79. Priest, N.K., Rudkin, J.K., Feil, E.J., Van Den Elsen, J.M.H., Cheung, A., Peacock, S.J., Laabei, M., Lucks, D.A., Recker, M. & Massey, R.C. From genotype to phenotype: can systems biology be used to predict *Staphylococcus aureus* virulence? *Nature Reviews Microbiology* **10**, 791-797 (2012).
80. Lynch, J.P.I. & Zhanel, G.G. *Streptococcus pneumoniae*: epidemiology and risk factors, evolution of antimicrobial resistance, and impact of vaccines. *Current Opinion in Pulmonary Medicine* **16**, 217-225 (2010).
81. Ramirez, M., Severina, E. & Tomasz, A. A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J Bacteriol* **181**, 3618-25 (1999).
82. Flores, C.O., Meyer, J.R., Valverde, S., Farr, L. & Weitz, J.S. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* **108**, E288-E297 (2011).
83. Bensing, B.A., Rubens, C.E. & Sullam, P.M. Genetic Loci of *Streptococcus mitis* That Mediate Binding to Human Platelets. *Infection and Immunity* **69**, 1373-1380 (2001).
84. Bensing, B.A., Siboo, I.R. & Sullam, P.M. Proteins PblA and PblB of *Streptococcus mitis*, Which Promote Binding to Human Platelets, Are Encoded within a Lysogenic Bacteriophage. *Infection and Immunity* **69**, 6186-6192 (2001).
85. Hsieh, Y.C., Lin, T.L., Lin, C.M. & Wang, J.T. Identification of PblB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone. *Sci Rep* **5**, 12265 (2015).
86. DeBardeleben, H.K., Lysenko, E.S., Dalia, A.B. & Weiser, J.N. Tolerance of a Phage Element by *Streptococcus pneumoniae* Leads to a Fitness Defect during Colonization. *Journal of Bacteriology* **196**, 2670-2680 (2014).
87. Harvey, R.M., Trappetti, C., Mahdi, L.K., Wang, H., McAllister, L.J., Scalvini, A., Paton, A.W. & Paton, J.C. The Variable Region of Pneumococcal Pathogenicity Island 1 Is Responsible for Unusually High Virulence of a Serotype 1 Isolate. *Infection and Immunity* **84**, 822-832 (2016).
88. de Stoppelaar, S.F., van 't Veer, C. & van der Poll, T. The role of platelets in sepsis. *Thromb Haemost* **112**, 666-77 (2014).
89. Cangemi, R., Casciaro, M., Rossi, E., Calvieri, C., Bucci, T., Calabrese, C.M., Taliani, G., Falcone, M., Palange, P., Bertazzoni, G., Farcomeni, A., Grieco, S., Pignatelli, P. & Violi, F.

Platelet activation is associated with myocardial infarction in patients with pneumonia. *J Am Coll Cardiol* **64,** 1917-25 (2014).
90. Rondina, M.T., Weyrich, A.S. & Zimmerman, G.A. Platelets as Cellular Effectors of Inflammation in Vascular Diseases. *Circulation Research* **112,** 1506-1519 (2013).
91. Loman, N.J. & Pallen, M.J. Twenty years of bacterial genome sequencing. *Nat Rev Micro* **13,** 787-794 (2015).
92. Harris, S.R., Cartwright, E.J.P., Török, M.E., Holden, M.T.G., Brown, N.M., Ogilvy-Stuart, A.L., Ellington, M.J., Quail, M.A., Bentley, S.D., Parkhill, J. & Peacock, S.J. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet Infectious Diseases* **13,** 130-136 (2013).
93. Doyle, M., Gasser, R., Woodcroft, B., Hall, R. & Ralph, S. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11,** 222 (2010).
94. Caskey, C.T. Using Genetic Diagnosis to Determine Individual Therapeutic Utility. *Annual Review of Medicine* **61,** 1-15 (2010).
95. Torok, M.E., Harris, S.R., Cartwright, E.J., Raven, K.E., Brown, N.M., Allison, M.E., Greaves, D., Quail, M.A., Limmathurotsakul, D., Holden, M.T., Parkhill, J. & Peacock, S.J. Zero tolerance for healthcare-associated MRSA bacteraemia: is it realistic? *J Antimicrob Chemother* **69,** 2238-45 (2014).
96. GPS partners. Global Pneumococcal Sequencing Project. Vol. 2016 (2016).

Addendum

Dutch summary (Samenvatting)
Author affiliation
Acknowledgement
List of publications
Other activities
Curriculum vitae

Dutch summary (Samenvatting)

Streptococcus pneumoniae, ook wel de pneumokok genoemd, is een Gram positieve bacterie met een ronde (-coccus) vorm die zich het beste thuis voelt in de neusholte (nasopharynx) van gezonde mensen. In sommige gevallen kan de bacterie infiltreren in plekken die normaal gesproken steriel zijn zoals de longen, het midden-oor, het bloed en de hersenen. Hier veroorzaakt de bacterie ernstige ziektes zoals longontsteking, oorontsteking en hersenvliesontsteking. De pneumokok is daarom een van de belangrijkste oorzaken van ziekenhuis opname en sterfte, meer dan 1 miljoen mensen overlijden aan infecties van deze bacterie per jaar. De belasting die invasieve pneumokokken infecties op de wereld gezondheid veroorzaakt behoeft onze aandacht.

Onze kennis van genoom sequencing om het genetische materiaal van een organisme in kaart te brengen en de toepassingen hiervan in de microbiologie zijn sterk toegenomen in de afgelopen jaren. Het doel van dit proefschrift was om deze sequencing technologieën te combineren met bioinformatische analyses en lab experimenten om zo ons begrip over *S. pneumoniae* te verbeteren. Dit werk heeft geresulteerd in uitgebreide studies van de genoom sequenties van honderden pneumokokken, geïsoleerd in de omgeving van Nijmegen over een periode van 10 jaar. Tezamen met de klinische informatie over ziekten veroorzaakt door deze bacterie en de vaccinatie status van de bevolking is onderzoek gedaan naar de opmaak van de genetische structuur van de populatie van pneumokokken en het effect op ziekte. Met deze informatie hebben wij gekeken naar hoe de genetische eigenschappen van de pneumokok ervoor zorgen dat deze ziekte wordt veroorzaakt, hoe deze ziekte zich manifesteert en hoe antibiotica resistentie zich ontwikkeld en verspreid. Verder is onderzocht hoe we nieuwe aangrijpingspunten voor antibiotica ontwikkeling kunnen ontdekken voor behandeling van ziekte veroorzaakt door de pneumokok.

Identificatie van nieuwe antibiotica

Resistentie tegen antibiotica stijgt aldoor en bedreigt de meeste, zo dan niet alle, eerstelijns antibiotica, inclusief penicilline. Als hier niets aan gedaan wordt, zullen simpele, behandelbare infecties niet meer te stoppen zijn. Om antibiotica resistentie bij invasieve pneumokokken infecties effectief te voorkomen zijn nieuwe antibiotica nodig. Voor het vinden van nieuwe antibiotica zullen nieuwe aangrijpingspunten geïdentificeerd moeten worden die essentieel zijn voor de groei en levensvatbaarheid van de pneumokok. In **hoofdstuk 2** wordt het gebruik van high throughput technieken en computationele benaderingen beschreven als alternatief voor tijdrovende en kostbare laboratorium technieken. Next generation sequencing methoden zouden een toevoeging of alternatief kunnen zijn voor traditionele kweek gebaseerde methodes om essentiële genen te vinden die gebruikt kunnen worden als aangrijpingspunten voor de ontwikkeling van nieuwe antibiotica. In **hoofdstuk 3** wordt een concept studie beschreven waarin we essentiële genen in pneumokokken ontdekken die als aangrijpingspunt voor nieuwe antibiotica kunnen dienen. Normaal gesproken is het ontwikkelen van nieuwe antibiotica

erg tijdrovend en duur, maar door slimme combinaties van genoom sequencing, bioinformatica en traditionele technieken hebben we 249 potentiele aangrijpingspunten ontdekt. Hieronder zitten 67 bekende eiwitten en hun producten, waarop de 75 FDA (Amerikaanse food and drug agency)-toegekende antibiotica voor zijn ontwikkeld. De nieuw gevonden eiwitten zouden kunnen worden als aangrijpingspunt voor ontwikkeling van nieuwe antibiotica. We hebben vier van deze nieuwe aangrijpingspunten experimenteel gevalideerd door gebruik te maken van commercieel beschikbare kleine moleculen die zouden moeten binden aan de eiwitten gevonden in deze studie. Deze methode zou antibiotica ontwikkeling in een stroomversnelling kunnen brengen.

Evolutie van de pneumokok tijdens een actief vaccinatiebeleid

Van de pneumokok is bekend dat deze zeer gemakkelijk DNA kan opnemen van uit zijn omgeving. Hierdoor kan hij snel nieuwe eigenschappen verkrijgen die helpen bij zijn overleving. Voor de behandeling van invasieve pneumokokken infectie zou meer kennis beschikbaar moeten zijn over hoe de pneumokok zich handhaaft tijdens interventies zoals vaccinatie. In **hoofdstuk 4** hebben we de genoom sequenties van 350 pneumokokken, geïsoleerd van volwassen patiënten uit 2 ziekenhuizen in Nijmegen, onderzocht om de genetische verschillen in de pneumokokken populatie te onderzoeken voor en na de invoering van een landelijk pediatrisch pneumokokken vaccin. We observeerden dat stammen met dezelfde genetische achtergrond vaak hetzelfde kapsel (suikers die een soort 'schild' vormen om pneumokok cellen) hebben. Ook zagen we dat ziekte veroorzakende stammen die een kapsel hadden waartegen gevaccineerd wordt verdwenen na introductie van het vaccin. Deze vermindering van ziekte bij volwassenen door stammen met dit kapsel suggereert dat er zogenaamde kudde immuniteit optreed door vaccinatie van kinderen. Het totaal aantal ziekte gevallen bij volwassenen bleef echter gelijk. Dit kwam door de toename van pneumokokken met een kapsel type dat niet in het vaccin zat. Twee mechanismen zijn hiervoor mogelijk: 1. verandering van kapsel van de bestaande pneumokokken populatie door opname van DNA of 2: toename van andere pneumokokken met een ander kapsel type. Wij observeerden bijna alleen situatie 2, toename van andere pneumokokken stammen. Genetische modificatie van het kapsel van de bestaande pneumokokken populatie werd niet veroorzaakt door het vaccin. We zagen echter wel een zeer drastische daling van de genetische diversiteit 1 jaar na introductie van het vaccin. Deze diversiteit keerde echter zeer snel terug naar een equilibrium in ziekte veroorzakende pneumokokken. Deze herordening van de diversiteit zou het wellicht mogelijk maken dat andere ziekte versterkende genen zich nestelen in de nieuwe populatie van pneumokokken, hiervoor zijn echter meer studies nodig. Dit onderzoek heeft ons begrip vergroot over hoe vaccinatie effect heeft op de pneumokokken populatie en laat zien dat genoom sequencen een belangrijke surveillance methode zou kunnen zijn om deze effecten in kaart te brengen hetgeen belangrijk is als nieuwe vaccins worden toegepast.

De relatie tussen genetische blauwdruk van de pneumokok, antibiotica resistentie en manifestatie van ziekte

In de toekomst zou het mogelijk kunnen zijn om zeer snel diagnostiek te kunnen toepassen op microbiële infecties door middel van genoom sequencing. Huidige fenotypische testen op antibiotica resistentie kosten enkele dagen, wat wellicht te lang is voor een adequate behandeling. Genotypische testen, waarbij het genoom van de infectie veroorzakende bacterie wordt gebruikt om te voorspellen welke resistentie deze heeft zijn nog niet uitontwikkeld. In **hoofdstukken 5 en 6** wordt het gebruik van genoom sequencing behandelt om inzichten te krijgen in welke genetische factoren de pneumokok gebruikt om resistent te worden en waarom bepaalde infecties een ernstiger verloop hebben dan anderen. In **hoofdstuk 5** hebben we de optelsom van de mutaties bepaald die nodig zijn voor volledige resistenties aan de hand van 1682 genoom sequenties van pneumokokken. We hebben mutaties gevonden die betrokken zijn bij resistenties tegen penicilline, trimethoprim, co-trimoxazole, erythromycine, ofloxacine, ciprofloxacine, and tobramycine. Resistentie werd veroorzaakt door mutaties in sleutel genen die betrokken zijn bij essentiële cellulaire processen. Ook vonden we enkele genen die, mits aanwezig, een verhoogde antibiotica resistentie geven. We laten zien dat een catalogus van deze resistentie veroorzakende mutaties en genen gebruikt kan worden om de mate van resistentie te voorspellen in meerdere pneumokokken. De studie in hoofdstuk 5 laat zien dat genoom sequencing gebruikt zou kunnen worden om nieuwe diagnostische testen te ontwikkelen waarbij snelle detectie van antibiotica resistentie centraal staat.

Invasieve pneumokokken infectie uit zich op veel verschillende manieren. De relatie tussen de genetische blauwdruk van de pneumokok en het ziektebeeld is een logisch verband om te onderzoeken gezien de diversiteit in het kapsel en de virulentie factoren in de pneumokokken populatie. Als er inderdaad genen zijn die betrokken zijn bij het verloop van de ziekte, zou informatie over de prevalentie in de pneumokokken populatie van belang zijn. In **hoofdstuk 6** worden alle genen van 350 invasieve pneumokokken onderzocht en welke relatie zij hebben met het veroorzaken van longontsteking, hersenvlies ontsteking en sterfte binnen 30 dagen na ziekenhuisopname. Onze bevindingen suggereren dat bepaalde genen inderdaad geassocieerd zijn met het ziekte beeld. Het moleculaire mechanisme hiervan is echter onbekend. In hoofdstuk 7 behandelen wij één van deze associaties van een pneumokokken gen dat mogelijk betrokken is bij binding aan bloedplaatjes. Dit gen bevindt zich op een bacteriofaag ingenesteld in het genoom van de pneumokok. We observeerden dat mensen die geïnfecteerd zijn door pneumokokken met dit gen een sterk verhoogd risico op sterfte binnen 30 dagen hebben. Interessant genoeg blijkt de bacteriofaag, met het gen dat betrokken is bij betere binding van de pneumokok aan bloedplaatjes, ingeschakeld te worden bij bepaalde antibiotica behandelingen. We observeerden dat de pneumokokken veel meer van het bloedplaatjes bindende eiwit tot expressie brachten na blootstelling aan fluoroquinolonen, een bepaalde klasse antibiotica. Evenwel zagen we geen verschil in

sterfte zelf na behandeling met deze antibiotica. Dit kan echter komen doordat het cohort van patiënten te klein is om deze verschillen op te merken. Over-activatie van bloedplaatjes gebeurt in patiënten met longontsteking en sepsis. In het verleden is aangetoond dat dit orgaan falen veroorzaakt, hetgeen veelal een doodsoorzaak is in patiënten die binnen 30 dagen overlijden. Identificatie van pneumokokken die een dergelijk gen bij zich dragen zou kunnen bijdragen aan het kiezen van een adequate behandelingsmethode.

Next generation sequencing en verdere bioinformatische analyses zijn technieken die uitermate geschikt zijn om de kennis over de pneumokok uit te breiden. Dit proefschrift benadrukt de noodzaak om dit soort nieuwe technieken te ontwikkelen om ziekte veroorzaakt door de pneumokok in kaart te brengen en in te dammen.

Author affiliation

Aldert L. Zomer[1,2,11], Amelieke J.H. Cremers[1,4], Andre J. van der Ven[3], Christa E. Gaast de Jongh[1], Elles Simonetti[1], Fredrick M. Mobegi[1,2,12], Gerben Ferwerda[1], Hester J. Bootsma[1,13], Jacques F. Meis[4,5], Jeroen Langereis[1], Marien I. de Jonge[1], Nel Roeleveld[6,7], Peter Burghout[1,14], Peter W.M. Hermans[1,15], Quirijn de Mast[3], Rahajeng N. Tunjungputri[3,8], Sacha A.F.T. van Hijum[2,9], Stefan P.W. de Vries[1, 16], Stephen D. Bentley[10].

[1] Laboratory of Pediatric Infectious Diseases, Department of Pediatrics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands.
[2] Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.
[3] Department of Internal Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.
[4] Department of Medical Microbiology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.
[5] Canisius-Wilhelmina Hospital, Department of Medical Microbiology and Infectious Diseases, Nijmegen, The Netherlands.
[6] Department for Health Evidence, Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands.
[7] Department of Pediatrics, Radboud Amalia Children's Hospital, Radboud University Medical Center, Nijmegen, The Netherlands.
[8] Center for Tropical and Infectious Diseases, Faculty of Medicine Diponegoro University - Dr. Karzai Hospital, Semarang, Indonesia.
[9] NIZO food research, Ede 6710 BA, The Netherlands
[10] Pathogen Genomics Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton Cambridge, United Kingdom.

Current address

[11] Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands.
[12] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[13] National Institute of Public Health and the Environment (RIVM), Beethoven, The Netherlands.
[14] Crucell-Johnson and Johnson, Janssen Pharmaceutical Companies of Johnson & Johnson, Leiden, The Netherlands
[15] Janssen Research and Development, Janssen Pharmaceutical Companies of Johnson & Johnson, Beerse, Belgium.
[16] Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

A

Acknowledgement

The hardest arithmetic to master is that which enables us to count our blessings. ~ Eric Hoffer (1898 – 1983).

A PhD journey is an emotional cocktail of exhilaration, exhaustion, victory, frustration, fright and euphoria. Overall, it has been a delightful four years of research and lots of experiences in this great institution, Radboud University Medical Centre; RUMC. This thesis appears in its current form due to the assistance and guidance of several people and organizations. It is my pleasure therefore to show my sincere gratitude by giving a word of appreciation in this thesis. Undoubtedly, there is likely to be a couple of omissions, for which I sincerely apologize.

First and foremost, I must express my most humble gratitude to my esteemed promotors Prof. Peter Hermans and Prof. Ronald de Groot for granting me an opportunity to undertake a Ph.D. research at the Laboratory of Pediatric Infectious Diseases. I will always value your warm encouragement, and thoughtful guidance during my research. I would also like to thank my thesis committee members: Prof. Heiman Wertheim, Prof. Martijn Huijnen, and Prof. Jan Maarten van Dijl for dedicating their time to read this thesis and provide critical guidance that led to this accomplishment.

Many thanks to my daily supervisors and co-promoters, Dr. Aldert Zomer and Dr. Sacha van Hijum, who were very committed in helping me throughout this research work. Being a supervisor is a privilege but it is frightening. I was privileged to have the two of you as my daily guides. You understood not only my professional life but also my personal one. Thank you for the trust, the insightful discussions, and the invaluable advice you offered. You continually offered critical comments and correction to the manuscripts that form this thesis. Dear Aldert, I appreciate your patience and guidance during my relocation period. You took your time to help me move my belongings to Nijmegen. All those years we shared an office, I always interrupted your thoughts to seek guidance or marvel over some amazing results. Your spirit of guidance and support will always go with me wherever my career takes me.

I am grateful to all my colleagues at the Laboratory of Pediatric Infectious Diseases (LPID); and the Centre for Molecular and Biomolecular Informatics (CMBI) for offering me a conducive working environment during my research. Thanks for your friendships, scientific advice and out-of-office adventures. I wouldn't forget Dr. Marien de Jonge (head of LPID): many thanks for your makeshift role as my supervisor during the last year of my PhD study. I remember the various collaborations you initiated, especially with the department of internal medicine and the thoughtful discussions towards shaping and improving the manuscripts that are included in this thesis. You also generously approved all my financial requests without constraint. I wouldn't particularly forget your efforts in assigning my wet-lab experiments to the technicians. With your help, we also secured a grant for me to visit and work at the Wellcome Trust Sanger Institute (WTSI). With your

support, I feel a much better and experienced scientist leaving than I was when I joined the LPID.

During my study, I was privileged to undertake part of my study at the Wellcome Trust Sanger Institute in Hinxton-Cambridge, the United Kingdom. My most humble gratitude to Prof. Stephen Bentley for having faith to collaborate in our research. I will be forever indebted to your support and scientific guidance during and beyond the visiting period. I thank the ever welcoming staff members who made my stay in Cambridge fruitful and memorable.

My accomplishments wouldn't be possible without the unconditional love and support from my family and friends. Many thanks to the Kenyan community in the Netherlands for spicing up my stay and making my free time adventurous. I will forever be grateful for your friendship and support. Am thankful for my new family: Mr. Samuel Ichangai and Mrs. Nancy Kirika, I wouldn't have prayed for better parents-in-law. You unconditionally welcomed me into your family and trusted me with your daughter. Thanks for your support, prayers, and guidance. Dear Judith, Catherine, Naomi, and Lorraine, I will always treasure you as my sisters; the fun, the intrigues, the chats. May God guide you as you strive to discover your paths in life.

Gratitude to my loving parents, Peter Moseti and Alice Mobegi. You rose above your humble backgrounds, adversities and career misfortunes, and sacrificed everything to ensure me and my siblings succeed in life. You guided me in times of difficulties and taught me family virtues. You raised me in the ways of the Lord; it is my hope that I turned out as you wished. Am thankful for all the invaluable life lessons you taught me about being independent, responsible, a hard worker, and open-minded. I thank my brothers Duncan Moseti, Tomkeen Onyambu, Erick Oseko, and Nicatious Manyange, and my sister Efrancia Kerubo. You were always there to cheer me up when I needed true friends. You never shied from reprimanding me and on your watch, I never lacked financially, socially, and spiritually.

Finally, I thank my wife and best friend Margaret Wambui for her immeasurable love, humor, compassion, and support during my doctoral training. You put up with my divided attention between studies and family time without complaint, and always listened tentatively and contributed to my research discussions at home. You are, and always will be the best part of my life; God bless your heart.

Above all, I wholeheartedly praise God the almighty for all the blessings. "I have fought a good fight, I have finished the race, and I have kept the faith". His grace granted me the capability to proceed successfully.

Fredrick

fredrickmaati@gmail.com

A

List of publications

Peer reviewed publications

1. **Mobegi, F. M**., A. Zomer, M. I. de Jonge, and S. A. van Hijum (2016). Advances and perspectives in computational prediction of microbial gene essentiality. Briefings Functional Genomics doi: 10.1093/bfgp/elv063.
2. **Mobegi, F. M**., S. A. van Hijum, P. Burghout, H. J. Bootsma, S. P. de Vries, C. E. van der Gaast-de Jongh, E. Simonetti, J. D. Langereis, P. W. Hermans, M. I. de Jonge and A. Zomer (2014). From microbial gene essentiality to novel antimicrobial drug targets. BMC Genomics 15(1): 958.
3. Cremers, A. J. H.*, **F. M. Mobegi***, M. I. de Jonge, S. A. F. T. van Hijum, J. F. Meis, P. W. M. Hermans, G. Ferwerda, S. D. Bentley and A. L. Zomer (2015). The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. Scientific Reports 5: 14952.
4. **Mobegi, F. M**., S. A. van Hijum, A.J.H. Cremers, S. D. Bentley, M. I. de Jonge and A. L. Zomer. Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data (Submitted).
5. **Mobegi, F. M**., A. J. H. Cremers, S. D. Bentley, M. I. de Jonge, S. A. van Hijum and A. Zomer. Genetic microbial correlates of the clinical manifestation of invasive pneumococcal disease (Manuscript in preparation).
6. Tunjungputri, R.N*, **F. M. Mobegi***, A. J. Cremers, C. E. van der Gaast – de Jongh, G. Ferwerda, J. F. Meis, Nel Roeleveld, S. D. Bentley, S. A. van Hijum, A. J. van der Ven, Q. de Mast, A. Zomer and M. I. de Jonge. Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. (Submitted).
7. de Vries S.P.W., S. Gupta, A. Baig, E. Wright, **F. M. Mobegi**, A. Wedley, L. LaCharme, A. N. Jensen, E. Pont, D. P. Wolanska, J. L'Heureux, T. Humphrey, P. Wigley, N. Williams, D. J. Maskell and A. J. Grant. Gene fitness analysis of the foodborne pathogen *Campylobacter jejuni* in host and environment (Manuscript in preparation).

***Authors contributed equally**

Abstracts and presentations

1. Mobegi, F.M., et al: Distance to antibiotic resistance: a bacterial genome-wide association study. CMBI Conference; 15th October 2015; Ravenstein, the Netherlands (Oral presentation).
2. Mobegi, F.M., et al: The post-vaccine microevolution of invasive pneumococci. 12th European pneumococcus meeting; 7th -10th July 2015; Oxford, United Kingdom (Oral presentation).
3. Mobegi, F.M., et al: The post-vaccine microevolution of invasive pneumococcal disease isolates. Bioinformatics & Systems Biology Conference; 20th -21st May 2015; Lunteren, the Netherlands (Oral presentation).
4. Mobegi, F.M., et al: The post-vaccine microevolution of *Streptococcus pneumoniae*. Netherlands Bioinformatics Conference; 7th -9th April 2014; Lunteren, the Netherlands (Poster presentation).
5. Mobegi, F.M., et al: From microbial gene essentiality to novel antimicrobial drug targets. Benelux Bioinformatics Conference; 9th -10th Dec 2013; Brussels, Belgium (Oral and poster presentation).
6. Mobegi, F.M., et al: From microbial gene essentiality to novel antimicrobial drug targets Netherlands Bioinformatics Conference; 15th-17th April 2013 (Oral and poster presentation).
7. Mobegi, F.M., et al: Rapid identification of potential drug targets for respiratory pathogens using Tn-seq and integrative genomics. Benelux Bioinformatics Conference; 10th-11th December 2012; Lunteren, the Netherlands (Poster presentation).
8. Mobegi, F.M., et al: Rapid identification of potential antimicrobial drug targets. ALW Platform for Molecular Genetics; 4th-5th October 2012; Lunteren, the Netherlands (Poster presentation).

Other activities

Main courses

Ph.D. course: Academic Writing. Radboud in 't to Languages, Radboud University Nijmegen 2012.

Awards and grants

Nijmegen Institute for Infection, Inflammation and Immunity (n4i) Travel grant 2013, Wellcome Trust Sanger Institute - Cambridge, United Kingdom.

Royal Netherlands Academy of Arts and Sciences; Academy Ter Meulen Grant 2014, Wellcome Trust Sanger Institute - Cambridge, United Kingdom.

Curriculum vitae

Fredrick Maati Mobegi was born in 1st October, 1984 in Kisii North district, Kenya. He obtained his primary leaving certificate in 1998 and a high school diploma in 2002. Thereafter, he joined Kenyatta University in Nairobi, Kenya to pursue a Bachelor of Sciences degree in Biochemistry and graduated (Hons) in 2008. Prior to his graduation, he took an internship with Kenya Tea Development Agency, where he worked on quality control in the processing of black tea. Fredrick's relentless quest for further education resulted in a Netherlands Fellowship Program (NFP) scholarship administered by NUFFIC in 2009 for a Masters course at Wageningen University and Research Centre (WUR), the Netherlands. He graduated, in 2011, with a Master of Sciences degree in Bioinformatics (MBF). His major thesis entitled "Genomic signatures of spore heat-resistance and growth temperature preference: a meta-analysis with *Firmicutes*" was completed under the auspices of Dr. Peter J. Schaap at the Laboratory of Systems and Synthetic Biology, WUR; and Dr. Lídia J. R. Lima at the Laboratory of Food Microbiology, WUR. His minor thesis entitled "Characterization of *Bovine* MHC Class I epitopes" was completed at the at Biosciences Eastern and Central Africa hub (BecA); International Livestock Research Institute (ILRI) in Nairobi, under the auspices of Dr. Peter J. Schaap (WUR), and Dr. Etienne de Villiers (BecA/ILRI). Afterwards, he completed a rotation with the University of California, Santa Cruz. In February 2012, Fredrick joined Radboud University Medical Centre as a PhD fellow in bioinformatics. He has developed scientific interest in medical bioinformatics, and he aim to translate basic bioinformatics tools into routine healthcare and biotechnology applications. Fredrick is currently as postdoctoral researcher at the Computational Cancer Biology group of the Netherlands Cancer Institute in Amsterdam.

Notes

Notes

Notes